НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»

Теплоенергетичний факультет

Кафедра автоматизації проектування енергетичних процесів і систем

До захисту допущено
Завідувач кафедри
_____ О.В. Коваль
(підпис) (ініціали, прізвище)
“___” _____2019р.

ДИПЛОМНА РОБОТА

на здобуття ступеня бакалавра

з напряму підготовки 6.050101 “ Комп’ютерні науки ”

на тему: «Система пошуку семантично-ідентичних фрагментів в різномовних текстах»

Виконав: студент IV курсу, групи ТР-51

Висоцький Володимир Іванович _____
(прізвище, ім’я, по батькові) (підпис)

Керівник доцент, доц., к.т.н. Лабжинський В. А. _____
(посада, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Консультант _____ _____ _____
(назва розділу) (вчені ступінь та звання, прізвище, ініціали) (підпис)

Рецензент Старший науковий співробітник інституту «ІПРІ НАН України», к.т.н. Сенченко В.Р. _____ _____
(посада, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2019 року

Національний технічний університет України
“Київський політехнічний інститут імені Ігоря Сікорського”

Факультет теплоенергетичний

Кафедра автоматизації проектування енергетичних процесів і систем

Рівень вищої освіти перший рівень

Напрям підготовки 6.050101 “ Комп’ютерні науки”

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ О.В. Коваль
(підпис)
”___”_____2019р.

ЗАВДАННЯ
на дипломну роботу студенту
Висоцькому Володимиру Івановичу
(прізвище, ім’я, по батькові)

1. Тема роботи: «Система пошуку семантично-ідентичних фрагментів в різномовних текстах»
керівник роботи Лабжинський Володимир Анатолійович, к.т.н., доцент
(прізвище, ім’я, по батькові науковий ступінь, вчене звання)

затверджена наказом вищого навчального закладу від ”__”___ 201__р. № __

2. Строк подання студентом роботи _____
3. Вихідні дані до роботи мова програмування Java, середовище IntelliJ IDEA, бібліотека Three.js.
_____
4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) розробити алгоритми пошуку семантично-ідентичних фрагментів в різномовних текстах, здійснити програмну реалізацію розроблених методів.
5. Перелік ілюстративного матеріалу графічне представлення інтерфейсу, приклади роботи програмного модулю_____
6. Дата видачі завдання ”10” жовтня 2019 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітки
1.	Вивчення та аналіз задачі	14.10.2018-23.12.2018	
2.	Розробка архітектури та загальної структури системи	2.02.2019-3.03.2019	
3.	Розробка структур окремих підсистем	4.03.2019-14.04.2019	
4.	Підготовка матеріалів	15.04.2019-18.04.2019	
5.	Програмна реалізація системи	18.04.2019-14.05.2019	
6.	Захист програмного продукту	15.05.2019	
7.	Оформлення пояснювальної записки	16.05.2019-3.06.2019	
8.	Передзахист	28.05.2019	
9.	Захист	17.06.2019-22.06.2019	

Студент _____ Висоцький В. І.
(підпис) (прізвище та ініціали,)

Керівник роботи _____ Лабжинський В. А.
(підпис) (прізвище та ініціали,)

АНОТАЦІЯ

Дипломну роботу виконано на 73 аркушах, вона містить 3 додатки та перелік посилань на використані джерела з 50 найменувань. У роботі наведено 6 рисунків.

Метою роботи було створення системи пошуку семантично-ідентичних фрагментів в різномовних текстах. Програма забезпечує пошук семантично-ідентичних фрагментів в різномовних текстах, забезпечує переклад текстів на англійську мову. Розроблений програмний продукт може бути використаний, наприклад, в організаціях та установах, де часто застосовується перевірка на плагіат.

ABSTRACT

The thesis is presented in 73 pages. It contains 3 appendixes and bibliography of 50 references. Six figures are given in the thesis.

The purpose of the work was to create a system for searching semantically identical fragments in multilingual texts. The program provides search of semantically identical fragments in multilingual texts, provides translation of texts into English. The developed software product can be used, for example, in organizations and institutions where plagiarization testing is often used.

ЗМІСТ

ВСТУП 6

1. ПОНЯТТЯ СИСТЕМИ ПОШУКУ СЕМАНТИЧНО-ІДЕНТИЧНИХ ФРАГМЕНТІВ В РІЗНОМОВНИХ ТЕКСТАХ 7

1.1 Огляд семантично-ідентичних фрагментів в різномовних текстах 7

1.2 Визначення семантично-ідентичних фрагментів в різномовних текстів 25

2. ПРОЕКТУВАННЯ СИСТЕМИ ПОШУКУ СЕМАНТИЧНО-ІДЕНТИЧНИХ ФРАГМЕНТІВ В РІЗНОМОВНИХ ТЕКСТАХ 46

2.1. Вибір і обґрунтування системи пошуку семантично-ідентичних фрагментів в різномовних текстах 46

2.2. Постановка задачі моделювання, обґрунтування припущень і розробку базової моделі, аналіз адекватності розроблених моделей.....47

2.3. Розробка алгоритму і методики проведення моделювання.....50

3. РЕАЛІЗАЦІЯ СИСТЕМИ ПОШУКУ СЕМАНТИЧНО-ІДЕНТИЧНИХ ФРАГМЕНТІВ В РІЗНОМОВНИХ ТЕКСТАХ.....53

3.1. Експериментальні дослідження системи визначення ідентичності різномовних текстів.....53

ВИСНОВКИ 55

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 57

ДОДАТОК 1 61

ДОДАТОК 2 63

ДОДАТОК 3 69

ВСТУП

Інструменти машинного перекладу (МП), що дозволяють працювати з текстами онлайн і здійснювати швидкий переклад, служать для інтернет-користувачів засобом комунікації.

Найчастіше результати роботи онлайн-інструментів вимагають постредагування і ефективно можуть використовуватися тільки тими, хто в якійсь мірі володіє мовою перекладу і мовою-джерелом.

Іншою проблемою є те, що не для всіх малих мов існують добре розроблені автоматичні перекладачі. Більшість систем при роботі з деякими парами мов використовують мову-посередник (зазвичай англійську мову). Інакше кажучи, переклад здійснюється не безпосередньо: спочатку відбувається трансфер тексту з мови-оригіналу англійською, а вже потім - на потрібну мову перекладу, що багато в чому впливає на якість перекладу.

Ми вибрали для роботи онлайн-перекладач Google Translate, по-перше, тому що на даний момент це найпопулярніший онлайн-перекладач у світі і ним користується більшість людей на нашій планеті, по-друге, тому що він перекладає майже з усіх мов світу і дуже легкий і зрозумілий у використанні.

Ми вважаємо що, ідентифікація основних проблем системи перекладу – це важливий крок у напрямку подальших досліджень.

Метою нашої роботи є пошук на плагіат семантично-ідентичних фрагментів в різномовних текстах та їх переклад за допомогою Google Translate з використанням усіх мов, які доступні.

Для досягнення поставленої мети нами були вирішені наступні завдання:

- вивчена історія розвитку систем МП, описані типи систем МП і розглянуті принципи їх роботи;
- досліджені поняття «якості перекладу» і способи оцінки якості перекладу;
- проаналізовані типи помилок, що з'являються при роботі системи ;

1. ПОНЯТТЯ СИСТЕМИ ПОШУКУ СЕМАНТИЧНО-ІДЕНТИЧНИХ ФРАГМЕНТІВ В РІЗНОМОВНИХ ТЕКСТАХ

1.1 Огляд семантично-ідентичних фрагментів в різномовних текстах

Передбачається, що інформація, яка міститься в текстових джерелах, може бути поданою різними мовами, що обумовлює необхідність перетворення різномовної вхідної інформації до єдиного її подання в базі знань. Таке подання інформації є основою для вирішення комплексу задач інформаційно-аналітичної діяльності (ІАД). Однак, зважаючи на велику різноманітність задач, які мають вирішуватися і вимагають специфічних методів їх автоматизації, це подання не для всіх задач є ефективним. В цьому зв'язку воно використовується безпосередньо лише для певного класу задач. Воно також є основою і для синтезу опису змісту, який відображається у формалізованому поданні. В інтересах вирішення таких задач ІАД, як: планування дій або прогнозування розвитку подій, укладається аксіоматична модель, яка містить відношення імплікативного характеру, або інша імплікативна система, яка обробляється логіко-математичними методами.

До формалізованого подання знань пред‘являються наступні вимоги [14]: по-перше, воно має бути представлено в такому вигляді, який забезпечить можливість коректної логіко-семантичної обробки знань; по-друге, воно має містити всю необхідну інформацію для вирішення конкретних інформаційно-аналітичних задач, тобто максимально повно зберігати текстове представлення елементів знань. З урахуванням цих вимог в якості формалізованого подання знань вибрана поняттєва структура (ПС) змісту природно-мовного тексту (ПМТ). Вона являє собою ієрархічну структуру, на верхньому рівні якої знаходяться найбільш загальні поняття і відношення між ними, кожний нижній рівень представляється поняттями і відношеннями, які конкретизують відповідні поняття і відношення найближчого вищого рівня. Іншими словами, верхній рівень ПС відповідає найбільш загальному

опису змісту тексту, нижчі її рівні відповідають рівням конкретизації цього опису. Кожне поняття і відношення в ПС супроводжується характеристиками, які відображають їхні властивості (понять і відношень), модальності та інші аспекти; лінгвістичною інформацією, яка характеризує мовні засоби їх відображення у вхідному тексті; семантичною інформацією, яка відбиває їх роль та інші характеристики (наприклад, об‘єкт, суб‘єкт, тип відношення, напрямок дії тощо). Сформована таким чином ПС містить всю необхідну інформацію для вирішення прикладних задач ІАД. Можливість її формування визначається наявністю відповідних знань в тезаурусі системи.

Особливості ПС полягають в наступному: її подання є гібридним і поєднує в собі властивості семантичних мереж і предикатних моделей (в якості вершин мережі виступають предикати); з метою уніфікації подання відношень, які в тексті можуть мати різну кількість аргументів, в ПС використовуються тільки одно- та двомісні предикати, для чого розроблено метод декомпозиції предикатів і предикатів вищих порядків на двомісні предикати першого порядку; для відображення рольових відношень введене поняття неявних предикатів; з метою зберігання виразових засобів природно-мовного текстового подання, введені спеціальні засоби - префікси і постфікси предикатів і понять, логіко-лінгвістичні зв‘язки, анафоричні посилання.

ПС, яка задовольняє сформульованим вимогам і містить всю необхідну інформацію як для подальшої її логіко-семантичної обробки, так і для синтезу опису ПС або її фрагментів природною мовою, формується в результаті лінгвістичної обробки ПМТ.

Слід відзначити, що при такому підході поняття, яке зустрілося на початку тексту і виявилося полісемічним, може уточнюватися наприкінці тексту. Крім того, при формуванні ПС тексту структурується його зміст, адже вся інформація, яка відноситься до одного поняття, де б вона не зустрілась в тексті, групується навколо цього поняття. Не має також принципового значення мова вхідного тексту, оскільки ПС його змісту не залежить від того, якими мовними засобами цей зміст

викладено. Це не стосується безпосередньо методів лінгвістичного аналізу, які саме і враховують закономірності виразу знань засобами певної мови.

Особливістю процесу синтезу опису ПС є то, що логічна структура синтезованого тексту визначається структурою ПС або вимогами користувача щодо змісту і обсягу вихідного текстового документу. В останньому випадку шляхом логічної обробки понятєвої структури "вилучаються" її необхідні фрагменти, що інтегруються в єдину ПС, яка є вже формалізованим поданням змісту тексту, який синтезується. Лінгвістичний синтез тексту здійснюється під управлінням структури і змісту елементів вхідної для нього ПС.

Отже, на першому етапі обробки інформації здійснюється приведення її до єдиної форми подання у вигляді понятєвої структури (для природно-мовних текстів шляхом вилучення знань з текстових джерел та їх формалізації). Природно-мовні тексти подаються англійською, російською та українською мовами. В базі знань інтегрується вся необхідна для комплексного аналізу апріорна ("стара") та поточна інформація. Методами логіко-семантичної обробки знань вся інформація аналізується на функціональну повноту, сумісність та протиріччя, під час чого виявляється також і хибна інформація. За допомогою цих же методів вирішуються і прикладні задачі, а саме узагальнення інформації, формування аналітичних оглядів та довідок тощо.

Ядром інструментальної знання-орієнтованої системи автоматизації обробки природно-мовної інформації є підсистема ТЕЗАУРУС [20]. Для комплексного вирішення задач автоматизації вилучення знань із природно-мовних текстів, їх формалізації і обробки в інтересах вирішення прикладних задач ТЕЗАУРУС містить три основних розділи: розділ, який містить лінгвістичні знання про мовні засоби тієї чи іншої природної мови; розділ, який містить загальні знання про реальний світ і знання з конкретних предметних областей; розділ, який включає знання про те, яким чином знання про реальний світ формулюються засобами конкретної мови. ТЕЗАУРУС, за своєю суттю є моделлю процесу відображення ("правил кодування") знань про світ конкретною природною мовою.

Підсистема ФОРМАЛІЗАТОР реалізує процедури вилучення і формалізації змісту (знань), відображеного в природно-мовних текстових джерелах. Підсистема УНІФІКАТОР реалізує процедури перетворення формалізованого подання знань до єдиного уніфікованого вигляду і може використовуватися самостійно. Однією з функцій цієї підсистеми може бути, наприклад, заміна таких відношень-дій, як “пересуватися”, летіти”, “йти”, “їхати” їх семантичним синонімом “рухатися”.

Внаслідок формалізації знань формується їх подання у вигляді поняттєвої структури. Сформована таким чином ПС містить всю необхідну інформацію для вирішення прикладних задач автоматизації інформаційно-аналітичної діяльності. Можливість її формування визначається наявністю відповідних знань у ТЕЗАУРУСІ.

Природно, що у відповідності зі сформованим таким чином формалізованим поданням змісту можна синтезувати і його опис природною мовою. Підсистему, яка реалізує цю процедуру, назвемо СИНТЕЗАТОР. Введемо також підсистему ІДЕНТИФІКАТОР, яка аналізує формалізовані подання знань чи їх фрагментів на ідентичність (тотожність) змісту, який вони відображають, і перетворюють ці подання до єдиного представлення.

Маючи в наявності розглянуті підсистеми, можна досить ефективно вирішити задачу ототожнювання фрагментів ПМТ на змістовому рівні і усунення дублювання текстових фрагментів з однаковим змістом. Крім того, якщо ФОРМАЛІЗАТОР може опрацьовувати різномовні текстові джерела (як-от: українсько-, російсько- та англомовні) і при цьому перетворює їх зміст до єдиного формалізованого подання, а СИНТЕЗАТОР може формувати по цьому поданню опис змісту, який воно відбиває, також різними мовами, то інформаційні системи, до складу яких надходять ці дві підсистеми, також є різномовними.

Комбінація ФОРМАЛІЗАТОР-СИНТЕЗАТОР дозволяє реалізувати машинний переклад і реферування текстових документів на основі аналізу саме їх змісту. Слід зазначити, що особливістю побудови такого перекладача є те, що переклад здійснюється не за окремими фразами (реченнями), як це робиться в

сучасних перекладачах, а спочатку вхідний текст повністю “опрацьовує” ФОРМАЛІЗАТОР, внаслідок чого формується уніфіковане формалізоване подання змісту тексту, а потім це подання “опрацьовує” СИНТЕЗАТОР. Під час формування формалізованого подання інформація, яка відноситься до деякого поняття і розосереджена по всьому тексту, концентрується навколо цього поняття. Таким чином здійснюється логічна структуризація змісту тексту. У випадку, коли ФОРМАЛІЗАТОР і СИНТЕЗАТОР функціонують у режимі однієї мови, це буде “машинний переказ” змісту документа, причому логічно структурований. Перевагою такого підходу до машинного перекладу є достатньо висока якість передачі змісту вхідного тексту.

В режимі реферування ПС тексту використовується частково, в залежності від потреб користувачів. Якщо користувачеві потрібен узагальнений реферат тексту, то в ПС виділяється прошарок лише верхнього рівня, який і сприймається в якості поняттєвого образу вхідного тексту. Правомірність використання терміну "образ" виправданий тим, що виділена частина ПС дійсно відбиває узагальнений зміст тексту. За виділеним поняттєвим образом синтезується його опис природною мовою, який і є узагальненим рефератом вхідного тексту. У випадку, коли користувач бажає отримати реферат з деяким рівнем деталізації тих аспектів змісту тексту, які його цікавлять, то з ПС крім верхнього прошарку виділяються відповідні фрагменти нижніх рівнів. За виділеними фрагментами ПС синтезується поняттєвий образ вхідного тексту, який і є основою для синтезу його реферату. Обсяг і зміст деталізації може задаватися користувачем шляхом переліку ключових слів або на основі природо-мовного запиту. Отже, розглядуваний підхід дозволяє формувати цілеспрямовані реферати, які відбивають потреби користувача в розширенні опису певних аспектів змісту тексту.

До викладеного слід додати, що розглянутий підхід дозволяє будувати різномовні системи автоматичного реферування, які здатні, наприклад, формувати реферати англо- і російськомовних текстів українською мовою.

Комбінація ФОРМАЛІЗАТОР-ІДЕНТИФІКАТОР-СИНТЕЗАТОР дозволяє автоматизувати вирішення задачі пошуку необхідної інформації за змістом запиту, сформульованого природною мовою.

Комбінація ФОРМАЛІЗАТОР-ІДЕНТИФІКАТОР з активним використанням ТЕЗАУРУСУ та УНІФІКАТОРА дозволяє автоматизувати вирішення наступних задач:

- усунення дублювання однакових за змістом різномовних текстових документів або їх фрагментів;
- автоматичне індексування різномовних текстових документів за їх змістом; вирішення цієї задачі також базується на формуванні формалізованого подання змістової сутності правил індексування документів;
- автоматичні класифікація та розподіл між тематичними рубриками текстової бази знань різномовних документів також саме за їх змістом.

Важливою задачею автоматизації ІАД є інтегрування знань в певній ПГ, які містяться в різномовних джерелах. Ця задача до теперішнього часу не ставилася навіть у постановочному плані. Для автоматизації цієї задачі введемо підсистему ІНТЕГРАТОР, функції якої проілюструємо на наступному прикладі. Нехай є англомовний текст, присвячений методам подання знань, у відповідності з яким ФОРМАЛІЗАТОРОМ сформована поняттєва структура його змісту. Таким же чином побудована поняттєва структура і за російськомовним текстом, в якому йдеться про методи обробки знань. Але в ньому також описані і методи подання знань, розглянуті в англомовному тексті. Тотожні за змістом фрагменти вхідних текстів породжують і тотожні фрагменти поняттєвої структури, оскільки їх подання уніфіковано як за формою, так і за мовою внутрішньо-машинного подання. Це дає можливість об’єднати поняттєві структури різномовних текстів з уніфікацією їх загальних частин. При необхідності можуть фіксуватися посилання на джерела, які породили ті чи інші фрагменти поняттєвої структури. Таким чином можуть накопичуватися знання в деякій предметній області, які містяться в різномовних текстових джерелах.

Однією з найважливіших задач обробки текстової інформації є визначення коректності, зокрема, логічної та змістової сумісності або суперечливості знань. Джерелами суперечностей можуть бути, наприклад, стилістичні огріхи у вхідному тексті, розпливчастість або недбалість формулювань, неповне інформування авторів, спотворення текстової інформації при її передачі по мережах, а також навмисне викривлення інформації (дезінформація). Автоматизація виявлення суперечностей в інформації на рівні обробки безпосередньо вхідного тексту – задача вкрай складна. Для автоматизації задачі визначення логічної і семантичної сумісності чи суперечливості знань, вилучених з ПМТ, введемо підсистему ЛОГІК. Функціонування цієї підсистеми базується на реалізації достатньо широкого спектру вже розроблених і розроблюваних сьогодні формальних методів логічного і семантичного аналізу знань на функціональну повноту, сумісність або суперечливість.

Основною задачею аналітичної діяльності є формування аналітичних оглядів і довідок. Знання-орієнтований підхід до розробки інформаційних систем дозволяє автоматизувати процес формування аналітичних оглядів і довідок у відповідності з вимогами користувача до їх обсягу і тематичної спрямованості, яка відбиває його інтереси. Це особливо важливо, якщо одні і ті ж джерела містять різнорідну і багатоаспектну в тематичному плані інформацію, а користувача цікавлять конкретні аспекти, наприклад, розвиток певних подій. При розглядуваному підході наявність підсистеми (назвемо її АНАЛІТИК), яка у відповідності із формалізованим поданням вимог до аналітичного огляду (довідки) здійснювала б пошук і локалізацію необхідних фрагментів знань у відповідній поняттєвій структурі, їх виділення і об’єднання в єдину, змістово цілісну структуру. Ця поняттєва структура і є основою для формування СИНТЕЗАТОРОМ тексту бажаного огляду (довідки) природною мовою. Для вирішення цієї задачі значна частина функцій АНАЛІТИКА може бути реалізована вже розглянутими раніше компонентами (ІДЕНТИФІКАТОР, УНІФІКАТОР, ІНТЕГРАТОР), що підкреслює їх уніфікованість відносно достатньо широкого кола прикладних задач.

Зіставне вивчення мов і їхніх систем набуває особливої актуальності в наш час, коли різнотипні міжмовні контакти і спілкування відіграють усе більшу роль у політичному, економічному та культурному житті переважної більшості країн і народів світу. З мовознавчої точки зору „дослідження споріднених і особливо близькоспоріднених, тобто таких мовних систем, які значною мірою збігаються, відкриває великі можливості для вдосконалення методів і прийомів лінгвістичного аналізу, відточує сам „інструментарій" пошуку, робить його тоншим і проникливішим"[15]. Важливим є і практичне вивчення зіставних досліджень не тільки для глибшого пізнання подібностей та відмінностей в лексиці й фразеології різних мов, але й для вдосконалення лексикографічних основ лінгводидактики. Набутий досвід зіставної лексикографії та лексикології органічно пов'язаний з лінгвокраїнознавством, етнолінгвістикою, лінгвокультурологією, теорією і практикою перекладу, сьогоднішніми проблемами міжкультурної комунікації, викликаними підвищеним бажанням глибшого оволодіння іноземними мовами. Мета статті – викласти у систематизованому вигляді основні принципи зіставного вивчення фразеології різних мов. Проблема „мова і культура" вже давно знаходиться в центрі уваги мовознавців. Тісний зв'язок мови й культури сьогодні не викликає сумніву через очевидний факт наявності в них рис, які об'єднують один народ із іншим і входять у компетенцію універсології як напряму в лінгвістиці, який вивчає загальні закономірності будови мов, єдність їхньої природи, структури й багатоманітності проявів. Виходячи з того, що різні мови – це не стільки різне позначення однієї і тієї ж речі, скільки різне її бачення, привернення уваги до різних проявів її буття, висновку про те, що мова наділена здатністю відображати світ у своїй неповторній семантичній системі своїм власним специфічним шляхом, будучи водночас універсальним засобом вираження мислення людини. За допомогою мови здійснюється особлива для кожного етносу форма передавання інформації, що охоплює норми, які є загальноприйнятними в тому чи іншому соціумі, його традиції, історію країни, життя і побут тощо. Мова збирає і зберігає скарби національної культури. При історичній зміні поколінь і суспільних формацій вона об'єднує народ у часі й просторі. Мова не існує поза культурою, яка глибоко пронизує повсякденне

життя народу, визначає його своєрідність і відрізняється в різних країнах певним національним колоритом.

Міжмовне спілкування й опанування іноземної мови неможливе без глибшого й докладного вивчення світу (не стільки мови, скільки світу) носіїв мови, їхньої культури в широкому етнографічному розумінні слова, їхнього способу життя, національного характеру, менталітету і т.ін., тому що реальне вживання слів у мові, реальна мовленнєва творчість значною мірою визначається знанням соціального й культурного життя колективу, який розмовляє тією чи іншою мовою. Актуальність цієї проблеми зростає у зв'язку зі зміною в галузі методики викладання іноземних мов, всередині якої спостерігається яскраво виражений поворот до тих чи інших явищ культури мови, що вивчається, переважно в її лексиці та фразеології – до проблематики етнолінгвістики й лінгвокультурології, яка формується на її основі [2]. Певні труднощі в цьому напрямі представляють і фразеологізми, що виражають національно забарвлене ставлення до світу, системи цінностей, способу життя, зберігаючи традиції тієї чи іншої етнокультурної спільноти. Через свою яскраву національну забарвленість вони завжди привертають особливу увагу вчених-лінгвістів, які вбачають у стійких словосполученнях певну зв'язну розповідь, викристалізований індивідуальний образ, що є міцним цементуючим чинником при взаємодії народів і культур. На думку Д.С.Лихачова, саме індивідуальні особливості народів пов'язують їх один з одним, змушують нас любити народ, до якого ми навіть не належимо, але з яким нас зіштовхнула доля. Отже, виявлення національних особливостей характеру, значення їх, роздуми над історичними обставинами, що сприяли утворенню їх, допомагають нам зрозуміти інші народи. Роздуми над цими національними особливостями мають важливе значення [13].

Особливі труднощі пов'язані з перекладом фразеосполучень, оскільки вони несуть у собі культурні й історичні цінності та сприймаються тільки через світорозуміння й світосприйняття носіїв певної культури, з чим часто пов'язане нерозуміння в них національної специфіки носієм іншої мови. Розумінню й перекладу фразеосполучень завдають труднощів також і лакуни. Міжмовна лакунарність є винятково цікавим явищем, що тісно пов'язане з проблемою концептів у

концептосфері мови. Ю.А.Сорокін дає таке визначення лакунам: „Лакунами називаються національно-специфічні елементи певної культури... , які з тих чи інших причин не передані в тексті рідною мовою або ті, які можуть стати перешкодою для розуміння тексту в культурі рідної мови” [22].

У простішому варіанті розглядає проблему лакунарності В.Г.Гак – як „пропуски в лексичній системі мови, відсутність слів, які, як видавалося, повинні були бути наявними в мові, якщо виходити з її відображальних функцій”[3]. В.І.Жельвіс лакунами називає те, що в одних мовах і культурах позначається як „окремість”, а в інших не сигналізується, тобто не знаходить суспільно закріпленого вираження [10].

Учені виділяють сьогодні такі типи лакун: предметні й абстрактні (за ступенем абстракції), родові й видові (за місцем у класифікаційних парадигмах), мотивовані й немотивовані (за позамовною зумовленістю), номінативні й стилістичні (за типом номінації), частиномовні (за належністю лакуни до певної частини мови).

В.М.Муравйов виділяє стилістичні лакуни на підставі відчутності в одній із мов слова (фразеологізму), що має те ж саме стилістичне забарвлення, як і слово з ідентичним значенням іншої мови [16]. Найкращим способом виявлення лакун у різних мовах є зіставний аналіз мов, що досліджуються.

Зіставлення фразеологічних одиниць, як нам видається, необхідно здійснювати комплексно, досліджуючи різні компоненти, що входять до їхнього складу: наочно-чуттєвий образ, емоціональний, культурний, конотативний тощо. У ході зіставного аналізу виявляються глибокі відмінності на різних рівнях фразеологічного концепту. Такий тип аналізу дозволяє виявити специфіку фразеологічних концептів різних мов, а також універсальні шляхи їхнього утворення на основі фразеологічної образності. Так, наприклад, при однакових базових лексемах фразеосполучень, що описують один і той самий наочно-чуттєвий образ і мають однакове значення на денотативному рівні, на рівнях фразеологізації в них спостерігаються суттєві відмінності.

Фразеологізми відображають багатий досвід народу, в них зберігаються уявлення про світ, у якому проживає той чи інший народ, його побут, культура. Саме у фразеологізмах добре репрезентована історія, ведення господарства, спосіб життя

людей будь-якої спільноти. Структура фразеологічної одиниці і значення, що закріплюються за ними як мовними знаками, мають глибоко національну структуру, яка завжди визначається умовами життя народу-носія певної мови, географічним середовищем, флорою і фауною, історією, суспільним устроєм, культурою, звичаями [9].

У вітчизняному й зарубіжному мовознавстві накопичений багатий досвід у цій галузі досліджень. Значний внесок у напрацювання проблем зіставного мовознавства зробили такі добре відомі вчені, як О.Потебня, І.О.Бодуен де Куртене, Є.Д.Поливанов, Г.О.Винокур, Л.В.Щерба, Г.Шухардт, Ф.Боас, Е.Сепір, Б.Уорф, Е.Бенвеніст. Велику роль у розвитку теорії і практики зіставного вивчення мов зіграли праці О.В.Федорова, який досліджував лексику на матеріалі російської, французької та німецької мов [26]. Неабиякий інтерес для теорії та практики зіставного вивчення германських мов у галузі лексики та фразеології представляють праці лейпцизької лінгвістичної школи, в яких зроблена спроба виявлення універсальних ознак лексики для її зіставлення з іншими мовами [28].

В.П.Конецька, аналізуючи праці германістів, підкреслює, що головними результатами зіставного дослідження повинні бути: 1) виділення універсалій, категорій і суттєвих ознак, релевантних для опису лексики; 2) обґрунтування одиниць, що порівнюються (порівняння на рівні словникового складу, мікросистем і слів як компонентів словникового складу на рівні елементарних значень як елементів семантичних полів); 3) установлення подібностей та відмінностей у певних сферах лексики конкретних мов [11].

Великого значення зіставному (контрастивному) аналізу мовних одиниць, зокрема фразеологічних одиниць надають З.Д.Попова та І.О.Стернін. На їхню думку, найбільш надійним способом виявлення національних особливостей концептів є зіставні дослідження, що дозволяють виявити наявність або відсутність, збіг або незбіг як самих концептів, так і мовних одиниць, які слугують для їх позначення [17]. На семантичному рівні виявлення специфіки окремих одиниць дає змогу простежити виникнення й розвиток дивергентних і конвергентних тенденцій, що проявляються у мовній свідомості народів, і може бути використано для моделювання концептів як

конституентів національної концептосфери. Крім того, специфічні семи, що виявилися у таких дослідженнях, інтерпретуються як відображення національно-специфічних ознак відповідних національних концептів і дозволяють моделювати концепт із вичленуванням його національно-специфічних ознак.

Національна своєрідність фразеології будь-якої мови є складним феноменом, що відображає як екстралінгвальні фактори, пов'язані з життям народу: особливості національного характеру, духовного складу, культури, своєрідність етнічного побуту, традицій, звичаїв, специфіку історичного розвитку народу, так і особливо лінгвістичні фактори, що визначаються специфікою лексико-семантичної і граматичної системи конкретної мови. Звідси стає очевидним той факт, що фразеологія є не тільки дзеркалом національної культури, вона є ще дзеркалом, у якому відображаються особливості системної організації мови, її будови. Сама історія розвитку лінгвістичної типології висуває на порядок денний необхідність у залученні фразеологічного матеріалу для створення найповнішої типологічної моделі мови. Включення фразеології в об'єкт лінгвістичного аналізу збагатить, на наш погляд, як саму теорію фразеології, так і загальне мовознавство.

Виходячи з викладеного вище, Д.О.Добровольський виділяє три різновиди фразеологічних універсалій: понятійно-фразеологічні універсалії (екстралінгвально зумовлені), лексико-фразеологічні та власне фразеологічні універсалії та дає їм докладну характеристику [4]. Здійснивши структурно-типологічний аналіз фразеології таких германських мов, як німецька, англійська та нідерландська, автор доходить висновку про те, що фразеологічна субсистема мови антропоцентрична за своєю суттю і, отже, всі структурно-типологічні закономірності у сфері фразеології опосередковані антропологічними факторами.

Лексичні розбіжності, пов'язані з неоднаковою продуктивністю лексем в утворенні фразеологічних одиниць проявляються у незбігу використання конкретних слів, що виступають компонентами фразеологізмів, які належать до одного й того самого семантичного поля. Ці відмінності пояснюються інертністю мовних традицій, неоднаковою символікою окремих представників, а інколи й цілих розрядів лексико-тематичних груп, особливостями семантичного членування понятійної картини світу.

Найближче до когнітивного дослідження фразеології підходить В.Т.Малигін, який робить висновок про те, що фразеологічні одиниці створюються певним соціумом і характеризуються його особливостями. Вони слугують чимось на зразок „ніші" для кумуляції світобачення і пов'язані матеріальною, соціальною або духовною культурою певної мовної спільноти, а тому можуть свідчити про її культурно-національний досвід і традиції [14]. Становлення й розвиток сучасних когнітивних досліджень дозволяє перейти до вивчення фразеосполучень із принципово нових позицій, вивчати їх не тільки як національно-культурну скарбницю, але і як джерело для розуміння когнітивних процесів, що ведуть до фразеологізації.

Когнітивний і лінгвокультурологічний напрями у вивченні фразеології, що дозволяють розглядати фразеологічні одиниці з точки зору їх національної та культурної специфіки, набувають за останній час усе більшого поширення [6]. Опора на принципи когнітології та лінгвокультурології дає змогу визначити місце й роль фразеологізмів у когнітивній базі того чи іншого народу, встановити їх значення як елементів мовної картини світу. За визнанням різних учених, ідіоматика, на відміну від слів, безпосередніше відображає когнітивну діяльність членів мовного колективу, що ґрунтується на наївному уявленні про світ носіїв певної мови, на їхньому ставленні один до одного, до того, що відбувається з ними у цьому світі. Навіть поверховий погляд на ідіоматику різних мов дозволяє зробити висновок про спільність багатьох „сюжетів", тобто подій та ситуацій, що знайшли своє відображення в ідіомах і про здатність кожної мови знайти виключно свої мовні засоби для позначення цих подій і ситуацій.

Фразеологізми, що відображають типові ситуації й уявлення, починають виконувати роль символів, еталонів, стереотипів культури. Таким чином, фразеологізми прямо (в денотаті) або опосередковано (через співвіднесеність асоціативно-образної основи з еталонами, символами, стереотипами національної культури) несуть у собі інформацію про навколишній світ і соціум, який у ньому проживає. Саме фразеологізми мовби нав'язують носіям мови особливе бачення світу й кожної в ньому ситуації. Особливого значення при цьому набуває розгляд

фразеологічних одиниць як знаків непрямої номінації, що несуть за собою інформацію про особливий, національний спосіб бачення світу тією чи іншою лінгвокультурною спільнотою.

Сьогодні наявні різні методики зіставного опису подібного й відмінного у фразеологічних системах тих чи інших мов. Так, Е.М.Солодухо говорить про міжмовні фразеологічні паралелі (ФП) [21], Д.О.Добровольський – про виділення конституентного, ситуаційного і фразеосемантичного інваріантів у семантиці фразеологічних одиниць [5], Ю.П.Сологуб – про виділення міжмовних фразеосемантичних еквівалентів (МФЕ) і міжмовних фразеосемантичних відповідників [20]. Оскільки для нас найбільш придатною є методика зіставного аналізу фразеологізмів, запропонована Ю.П.Сологубом, зупинимося на характеристиці виділених ним типів МФЕ і МФВ. Відповідно до його концепції, внаслідок глибокого аналізу семантики і структури стійких мовних зворотів можна виділити міжмовні фразеологічні еквіваленти (МФЕ), що відображають семантичну тотожність системи образів, і міжмовні фразеологічні відповідники (МФВ), що відображають лише близькість образної інтерпретації дійсності за допомогою фразеологічних засобів.

У міжмовних фразеологічних елементах (МФЕ) об’єднуються такі чотири типи ідіом:

1) МФЕ з повною однозначною відповідністю одиниць лексичного й граматичного плану, наприклад: англ. to set the wolf to keep the sheep – укр. поставити вовка овець стерегти – рос. поставить волка овец стеречь; англ. to clip smb’s wings – укр. підрізати крила кому-небудь – рос. подрезать крылья кому-либо; англ. to cross (або pass) the Rubicon - укр. перейти Рубікон, рос. перейти Рубикон.

2) МФЕ з відсутністю повної однозначної відповідності одиниць лексичного плану, наприклад: англ.bury (або hide) one’s head in the sand - укр. закривати на факти очі – рос. закрывать на факты глаза; англ. to measure another man’s foot by one’s own last – укр. судити про інших по собі – рос. судить о других по себе; англ. an April fool – укр. жертва першоквітневого жарту – рос. жертва первоапрельской шутки.

3) МФЕ з відсутністю однозначної відповідності одиниць граматичного рівня, наприклад: англ. lid is on – укр. інформація приховується, відомості не розголошуються – рос. информация утаивается, сведения не оглашаются;англ. a lie out of (the)whole cloth – укр. зухвала, крикуща брехня – рос. наглая, вопиющая ложь ; англ. next door – укр. поряд, недалеко, дуже близько, по сусідству – рос. рядом, недалеко, очень близко, по соседству.

4) МФЕ змішаного типу, наприклад: англ. dry as dust – укр. скучний, нецікавий, написаний сухою мовою – рос. скучный, неинтересный, написанный сухим языком; англ. as the crow flies – укр. найкоротшим шляхом, по прямій (лінії), навпростець; англ. in all intents and purposes – укр. фактично, насправді, по суті, дійсно – рос. фактически, на деле, по существу, в сущности, действительно.

У міжмовних фразеологічних відповідниках (МФВ) об'єднуються такі два типи ідіом:

1) МФВ I – ідіоми, які характеризуються подібністю образно-мотиваційних основ, що зумовлено однаковою інтерпретацією реалій навколишнього світу, наприклад: англ. for better (or) for worse; укр. у щасті і нещасті, на радість і горе - рос. – в счастье и несчастье, на радость и горе; англ. a peroxide blonde – укр. блондинка з пофарбованим волоссям, „хімічна" блондинка – рос. блондинка с крашенными волосами, „химическая блондинка".

2) МФВ II - ідіоми, в основі семантики яких лежить спільна логіко-семантична формула, реалізована у значеннях фразеологічних одиниць за допомогою різних образів, наприклад: англ. to show the elephant – укр. знайомитися з визначними місцями – рос. знакомиться с достопримечательностями; англ. no end of - укр. дуже багато, маса, безліч – рос. очень много, масса, уйма.

Якщо типи МФЕ відрізняє порівнянність як структурно-граматичної організації ідіом, так і їхньої семантики в різних мовах, то МФВ відображають лише подібні або спільні логічні підстави як семантичні основи фразеологічного образу. У вітчизняному мовознавстві зіставний опис фразеологічних систем проводиться з використанням традиційних термінів „план змісту" (семантика) фразеологічних одиниць і „план вираження" (лексичний склад компонентів і їхні граматичні

характеристики). У зарубіжній лінгвістиці часто використовуються терміни „глибинна й поверхова структура мовних одиниць”, у тому числі й фразеологічних. Спільність плану вираження фразеологізмів у різних мовах може бути виявлена в ході опису семантики й граматичної природи їх компонентів, а спільність плану змісту – в ході виявлення міжмовних паралелей шляхом фразотворчого моделювання й тематико-ідеографічної систематики фразеологізмів, що виявляє їх образно-мотиваційні основи. Зіставний аспект системного вивчення фразеології безперечно представляє великий інтерес із точки зору лінгвокультурології та етнолінгвістики в цілому. Кінцевою метою подібних досліджень є зіставлення цілих фразеологічних фондів мов, що досліджуються. Найтиповішими мікросистемами фразеологічних одиниць у мовах світу є мікросистеми, які виражають: емоційний стан людини; мікросистеми, що характеризують ті чи інші життєві ситуації; мікросистеми, для яких характерно відображення тих чи інших інтелектуальних якостей людини. Усі дослідники звертають увагу на те, що розподіл фразеологічних одиниць за семантичними розрядами нерівномірний, не всі сфери об’єктивної дійсності отримують відображення у фразеології. Одні семантичні групи охоплюють більшу кількість фразеологічних одиниць, інші – меншу, деякі представлені поодинокими зворотами. Наповнюваність тих чи інших груп знаходиться у прямій залежності від конотативного насичення фразеологізмів, що є прямим наслідком національного складу мислення. Аналізуючи свої спостереження над семантичними властивостями фразеологізмів російської та української мов, Н.Ф.Алефіренко виділяє дві великі групи одиниць: семантично адекватні фразеологізми і ті з них, які характеризуються відмінними семантичними рисами [1]. Фразеологізмам першої групи притаманна адекватна семантика, образи, що збігаються, і лексемний склад; ці одиниці виражають одні й ті ж поняття. До фразеологізмів другої групи, які розрізняються за семантикою, автором віднесені фразеологічні запозичення, утворені на ґрунті власне російської або української мови, які згодом поширилися у спорідненій мові. У переважній більшості випадків відмінності стосуються ступеня інтенсивності конотації фразеологічних одиниць другої виділеної групи. У статті також указуються генетичні основи формування семантики фразеологічних одиниць обох груп. Своєрідне

термінологічне позначення типів міжмовних відповідників ми знаходимо у дослідженні О.С.Шакірова. Семантично тотожні різномовні нумеративні фразеологічні одиниці визначаються ним як міжмовні універсалії незалежно від того, чи збігається їхній компонентний склад і образні структури. До нумеративних фразеологічних одиниць-еквівалентів автором відносяться точні кальки, тобто перекладені з одної мови іншою фразеологізми, лексико-граматичний склад яких відтворений без будь-яких змін. У своєму дослідженні Н.Ю.П'ятницька аналізує декілька типів співвідношень міжмовних еквівалентних фразеологічних одиниць: ті, які повністю збігаються за структурою й семантико-стилістичними властивостями, ті, які частково збігаються за структурою, але ідентичні за значенням і стилістичним забарвленням, різних за структурою, але ідентичних за семантико-стилістичними якостями [18]. Заслуговують на увагу спостереження авторки за впливом неяскравості фразеологічних образів і відсутності їх національного забарвлення на міжмовну еквівалентність фразеологізмів. Якщо різномовні фразеологізми, збігаючись за значенням, відрізняються національно-фразеологічною образністю, вони відносяться до міжмовних синонімів.

Для проведення зіставного аналізу фразеологічних одиниць російської та німецької мов М.Н.Сенченко визначає критерії тотожності та різних ступенів розбіжностей між фразеологічними одиницями двох мов [19]. До тотожних він відносить узуальні російські та німецькі одиниці, для яких характерний збіг денотативного й конотативного аспектів фразеологічного значення, структури й лексичного складу. Незбіг цих критеріїв у змісті й формі фразеологічних одиниць характеризує фразеологічну розбіжність. Залежно від того, чи стосуються незбіги одного, декількох або всіх критеріїв, виділяються три ступені розбіжностей. При повній розбіжності спостерігається семантичне зміщення й незбіг конотативного аспекту значення, структури й лексичного складу. Вдалою, на наш погляд, спробою є системний розгляд фразеології у зіставному плані М.П.Кочерганом [12]. З огляду на значення фразеологізмів, автор підходить до їх класифікації, розрізняючи такі типи міжмовних співвідношень, як повна еквівалентність, неповна еквівалентність із її підгрупами та безеквівалентність з двома можливими типами співвідношень.

Важливе місце в описі національної специфіки фразеологізмів у зіставному плані відводиться опису їхньої семантики й структури, в особливостях вживання та семантичних зв'язках у фразеологічних системах. Робиться цілком обґрунтований висновок про те, що національно-мовна специфіка фразеології найбільшою мірою виявляється в лексичному складі фразеологізмів, частотності певних тематичних груп слів, у варіативності фразем, їх багатозначності і системних (синонімічних й антонімічних) зв'язках [12].

Отже, здійснений нами короткий огляд основних принципів зіставного вивчення фразеології різних мов переконливо показав, що в основу визначення типів міжмовних фразеологічних відповідників/невідповідників, як правило, покладені збіги (незбіги) семантики, граматичної організації та компонентного (лексемного) складу різномовних фразеологічних одиниць. Однак план змісту характеризується ними по-різному: як сукупний зміст фразеологічних одиниць, значення, стилістичне забарвлення, фразеологічний образ, семантико-стилістичні властивості фразеологізмів і т.ін. Звідси випливає, що, зіставляючи фразеологічні одиниці двох і більше мов, ми в першу чергу повинні зіставляти їхнє значення, семний склад, сигніфікативно-денотативний і конотативний макрокомпоненти. Семантична тотожність або відмінність різномовних фразеологізмів означає тотожність або відмінність їхнього семного складу, у спрощеному вигляді набір мінімальних смислових компонентів сигніфікативно-денотативного й конотативного складників фразеологічного значення. Надання переваги семантичній тотожності/відмінності при виявленні типів міжмовних фразеологічних відповідників/невідповідників означає, що організуючою теорією при визначенні цих типів може слугувати компонентна теорія, в основі якої лежить метод компонентного аналізу. Такий підхід при вирішенні питання про критерії тотожності й відмінності між фразеологізмами зіставлюваних мов видається нам найбільш перспективним у подальших дослідженнях.

1.2 Визначення семантично-ідентичних фрагментів в різномовних текстів

У теорії перекладу еквівалентність визначають [15], як збереження у процесі перекладу відносної змістової, семантичної, стилістичної і функціонально-комунікативної рівності інформації, що міститься в оригіналі й перекладі. Еквівалентність перекладу залежить насамперед від ситуації породження тексту оригіналу і його відтворення в мові перекладу [там само, с.19].

Розрізняють [27] потенційно досяжну еквівалентність, де наявна максимальна спільність змісту двох різномовних текстів, що допускається через різницю мов, на яких створені ці тексти, і перекладацьку еквівалентність. При цьому межею перекладацької еквівалентності автор вважає максимально можливий (лінгвістичний) ступінь збереження змісту оригіналу при перекладі, але в кожному окремому перекладі змістова близькість до оригіналу різного ступеня й різними способами наближується до максимального.

Такі дослідники, як М. Брандес і В. Провоторов фактично замінюють еквівалентність тотожністю, стверджуючи, що переклад має повністю зберігати зміст оригіналу [10]. Тому цілком природно, що в теорії і практиці перекладу оперують такими подібними поняттями, як еквівалентність, адекватність і тотожність. У широкому значенні еквівалентність розуміється як щось рівноцінне, рівнозначне чому-небудь, адекватність тлумачиться як щось цілком рівне, а тотожність трактується як щось, що є повним збігом, подібністю з чимось. За твердженням російського вченого Г.Д. Воскобойника [16], тотожність виступає універсальним епістемологічним принципом перекладацької теорії і практики, та репрезентується двома різновидами – позитивістським та феноменологічним.

Тому саме найменша семантична категоричність слова «еквівалентність» і сприяла його найбільшій уживаності у сучасному перекладознавстві. Хоча, звичайно, поняття адекватності, тотожності, повноцінності і навіть аналогічності залишаються в тому самому семантичному полі, що й термін «еквівалентність» й іноді дублюють один одного. У трактуванні В.С. Виноградова під еквівалентністю, в теорії перекладу

слід розуміти збереження відносної рівності змістовної, смислової, семантичної, стилістичної й функціонально-комунікативної інформації, що міститься в оригіналі і перекладі. Тут автор особливо наголошує, що еквівалентність оригіналу і перекладу полягає насамперед у спільності розуміння інформації, що міститься в тексті, у тому числі й тієї, яка впливає не лише на розум, але і на почуття реципієнта і яка не тільки експліцитно виражена в тексті, але й імпліцитно співвіднесена до підтексту. Еквівалентність перекладу залежить також, на думку Виноградова [15], від ситуації породження тексту оригіналу і його відтворення в мові перекладу.

У свою чергу, Ю.А. Найда [59] виокремлює формальну і динамічну еквівалентність перекладу. Формальна еквівалентність полягає у прагненні перекладача відтворити повідомлення якомога ближче до форми й змісту оригіналу, передаючи максимально точно його загальну структуру й окремі її складові елементи.

Формальна еквівалентність полягає, за твердженням автора, у прагненні перекладача відтворити повідомлення наближено до форми й змісту оригіналу при цьому передаючи максимально точно його загальну структуру й окремі її складові елементи.

У випадку динамічної еквівалентності перекладач ставить за мету створення між текстом перекладу та його читачем такого самого зв’язку, який існував між оригінальним текстом і його читачем – носієм мови.

Окрім еквівалентності перекладачеві також потрібно розуміти поняття адекватності і тотожності перекладу.

Як відомо, переклад потребує концентрації і точної передачі змісту перекладеного матеріалу. У зв’язку з цим, необхідно з’ясувати поняття «адекватність» й «еквівалентність», які дещо відрізняються один від одного.

Адекватність трактується, як вичерпна передача смислового змісту оригіналу і повна функціонально-стилістична йому відповідність. Таким чином, адекватний переклад слід розуміти як відтворення єдності змісту і форми оригіналу засобами іншої мови.

За В.В. Балахтаром та К. С. Балахтар, переклад тексту можна вважати адекватним, якщо хоча б одна з двох нижчевказаних умов збережена [6]:

- правильно перекладені всі терміни та їхні сполучення;
- переклад є зрозумілим для спеціаліста, оскільки він чітко розуміє зміст написаного і не має до перекладача ніяких питань і зауважень.

Відповідно до теорії Н.В. Складчикової у перекладознавстві є необхідність виокремлювати чотири основні параметри адекватності перекладу [48]:

- параметр адекватності передачі семантичної інформації;
- параметр адекватності передачі емоційно-оцінної інформації;
- параметр адекватності передачі експресивної інформації;
- параметр адекватності передачі естетичної інформації.

Безперечно в поняття адекватність входить передача стилістичних й експресивних відтінків оригіналу. Крім того, навіть за відсутності формальної точності передачі окремих слів та словосполучень переклад у цілому може бути адекватним. Як зауважує Р.К. Міньяр–Білоручев [44]., переклад є адекватним саме завдяки порушенню цієї елементарної і поверхової точності. Він уважає що це відбувається, тоді коли окремі, другорядні елементи тексту передаються у повній відповідності із задумом автора, відтак переклад досягає високого ступеня адекватності [44].

У поняття адекватний переклад В.М. Комісарова [34] вкладає більш широкий смисл, а тому може використовуватися як синонім коректного перекладу, який забезпечує необхідну повноту міжмовної комунікації в конкретних умовах.

Обов'язковою умовою адекватного перекладу для лінгвістів В. І Карабан., О. В. Борисова, Б. М. Колодій, К. А. Кузьміна, є вміння правильно аналізувати граматичну будову іншомовних речень, мати широкі знання граматичних норм цільової мови та сталі навички мовлення мовою перекладу, у тому числі навички вживання у перекладі слів, словосполучень та синтаксичних конструкцій, які є у мові перекладу і які відсутні у мові оригіналу [27].

Поділяючи погляди вище згаданих науковців, А.В. Федоров [52] під адекватністю розуміє повноцінний переклад і визначає повноцінність перекладу як вичерпну передачу змісту оригіналу і повноцінну функціонально-стилістичну йому відповідність. Повноцінність перекладу, на його думку, полягає у передачі

специфічного для оригіналу відношення змісту і форми шляхом відтворення особливостей останньої чи створення функціональних відповідностей цим особливостям.

Таким чином, із розглянутого вище можна зробити висновок про те, що адекватний переклад є передачею сказаного автором зі збереженням точності формату тексту, мови, стилю, задуму автора та стилістичних відтінків оригіналу. У випадку, коли під час перекладу відбувається втрата формальної точності, переклад теж можна вважати адекватним.

Водночас буває і так, що переклад є адекватним саме завдяки порушенню поверхової точності. Головною рисою адекватного перекладу є збереження концепції вихідного тексту.

Цілком природним тому є порівняння роботи перекладача росіянином В.В. Левик [40], відомого як майстра практики і дослідника теорії перекладу з мистецтвом фокусника-іллюзіоніста, позаяк саме його робота ґрунтується на обмані. Перекладач створює зовсім інше, зовсім несхоже на оригінал, але обманює нас ілюзією повної подібності. Утім щоб остаточно зрозуміти принцип та суть еквівалентного перекладу, потрібно знати його визначення.

- Програми, які дозволяють конфігурацію правил корпоративної мови та контролюють дотримання технічним автором заданих правил у процесі продукування тексту, розробляються як універсальні продукти для конкретної мови, що можуть адаптуватись до потреб великого, середнього та малого бізнесу, а також використовуватись фрилансерами, які спеціалізуються в певній предметній галузі й орієнтовані на створення «індивідуальної» контрольованої мови. Для німецької мови розроблено два комерційні продукти такого типу:
- Acrolinx Suite, розробник – фірма Acrolinx GmbH, Берлін, ФРН;
- CLAT (Controlled Language Authoring Technology), розробник – Інститут спілки сприяння прикладним дослідженням у галузі інформації (Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes, IAI), Саарбрюкен, ФРН;

- продемонструємо функціональні можливості та обмеження програмних продуктів такого типу на прикладі німецькомовної версії програми CLAT, розробленої для роботи з англійською та німецькою мовами.
- CLAT складається з модулів: CLAT-сервер, UMMT (Utility for Mandate Management Tasks) та клієнти: Java CLAT-Client та CLAT-In для Word та CLAT-In для PTC Arbortext, архітектуру яких унаочнює рис. 1.1:

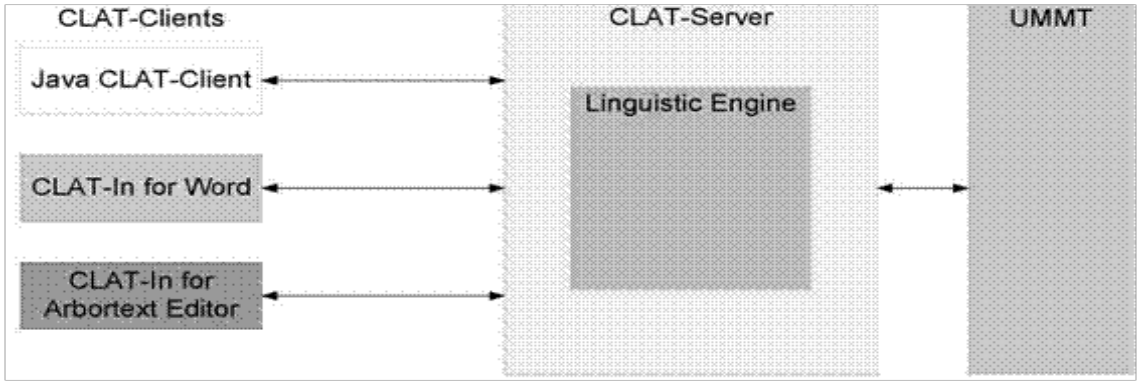

Рис. 1.1. Компоненти технології CLAT [484]

Основний модуль CLAT – UMMT, користуючись яким адміністратор створює нові проекти, генерує існуючі проекти і регламентує ролі та права користувачів. Для кожного проекту створюються специфічні ресурси (база термінів і словник користувача) та правила рестрикції тексту (за категоріями «термінологія», «правопис», «граматика», «стиль»), на ґрунті яких продукуються тексти мовою оригіналу.

Робочу область термінологічної бази в UMMT візуалізує Рис. 1.2.:

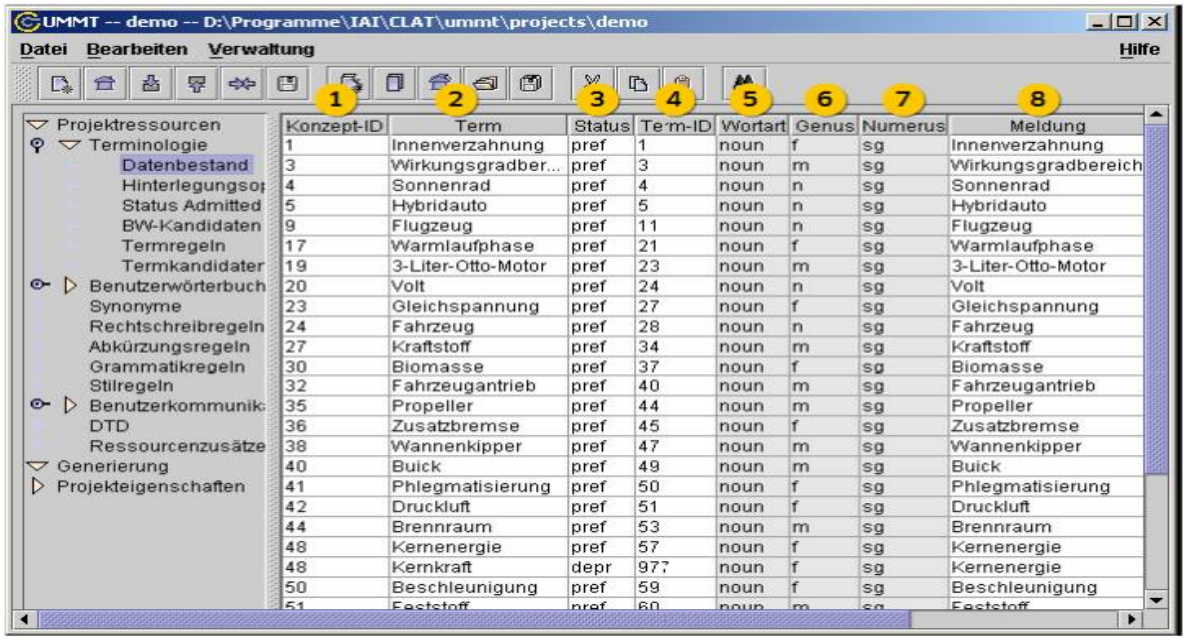

Рис. 1.2. Робоча область термінологічної бази в UMMT

Так, у термінологічну базу проекту вносяться: 1) *Konzept-ID* «ідентифікаційний номер концепту», який дозволяє пов'язувати синонімічні терміни зі статусом «preferred» – рекомендований, «admitted» – дозволений та «deprecated» – заборонений; ця функція використовується для контролю за вживанням рекомендованих термінів у текстах технічної документації, наприклад, якщо автор вживає в тексті термін зі статусом «дозволений» або «заборонений», CLAT-клієнт підказує «рекомендований» до вживання термін через ідентифікаційний номер концепту; 2) *Term* «термін» – термін; 3) *Status* «статус терміна»; 4) *Term-ID* «ідентифікаційний номер терміна»; 5) *Wortart* «частина мови»; 6) *Genus* «граматичний рід»; 7) *Numerus* «число»; 8) *Meldung* «повідомлення» – у це вікно вноситься термін із статусом «рекомендований», який виводиться програмою як підказка, якщо технічний автор уживає термін іншого статусу.

Через ресурс «Словник користувача» у термінологічну базу UMMT вносяться корпоративні терміни або прототерміни, що не зафіксовані у нормативних словниках, напр.: *BioGRAN*. Це виключає їхнє маркування як неправильних у процесі перевірки тексту програмою.

Фахові та корпоративні мови не виключають явища синонімії, тому для програми передбачено ресурс *Synonyme* «синоніми», у який вносяться абсолютні,

напр. *Bautyp* <–> *Bauart*, та ідеографічні синоніми, напр. *Trinkgefäß* –> *Trinkglas*. Тип синонімії маркується символами <–> та –>. Вікно роботи з синонімами ілюструє рис. 1.3:

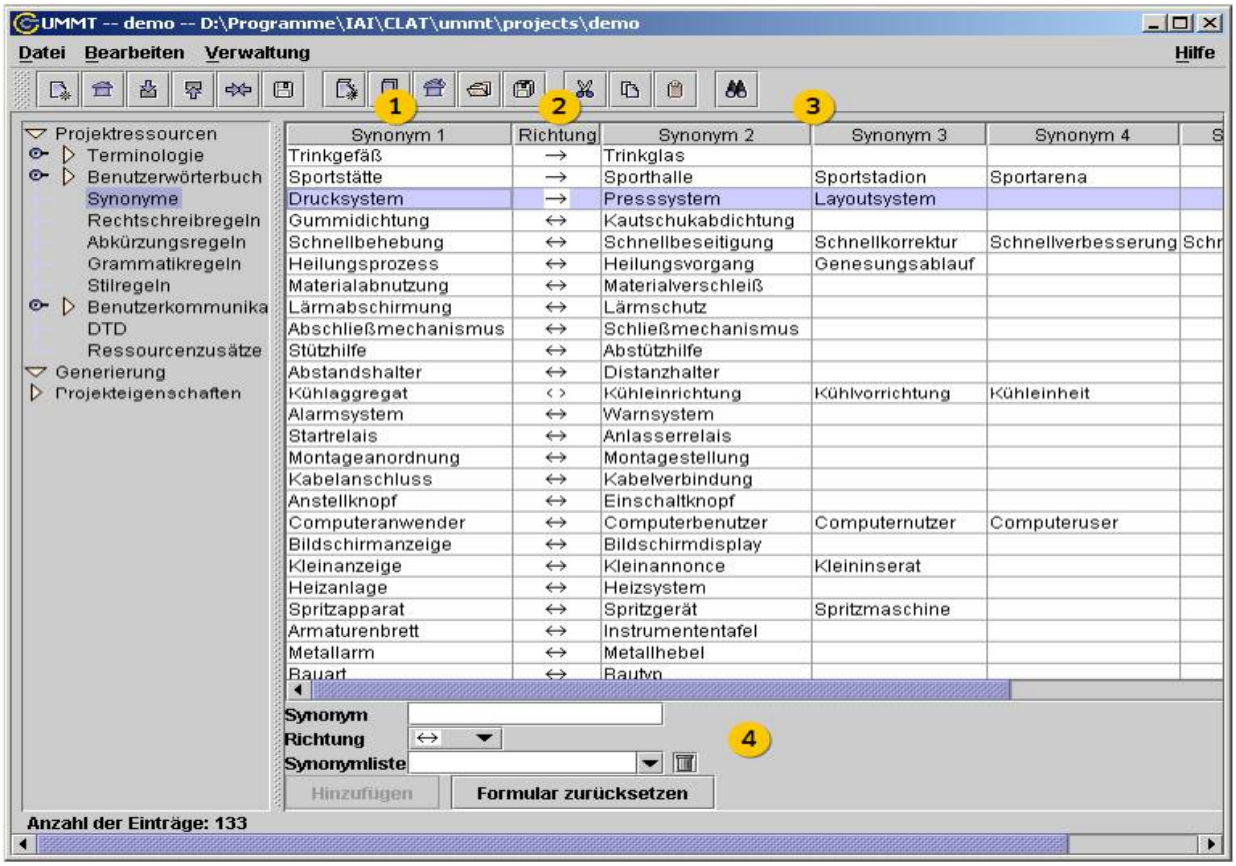

Рис. 1.3. Робоча область ресурсу «синоніми» в CLAT-UMMT

Додаткова функція CLAT – автоматична екстракція «кандидатів у терміни», які можна надсилати термінологам для аналізу або додавати до термінологічної бази проекту.

Функції контролю в CLAT реалізуються на ґрунті синтаксичного та морфологічного аналізу і синтезу. Межі слів розпізнаються програмою за знаками пунктуації або пробілами, а межі речень, якщо це не структурована мова файла (SGML / XML), – за знаками пунктуації, пробілами та написанням першого слова з великої літери. Кожне слово аналізується як лінгвістична категорія. В процесі аналізу слово розкладається на морфеми, кожна з яких порівнюється з морфемами морфологічного словника німецької мови, ідентифікується статус морфеми (коренева морфема, словотворча морфема, флексія), аналізується сполучуваність з іншими

морфемами. На другому етапі ідентифіковані морфеми синтезуються у слова згідно з правилами німецького словотвору. Багатозначність морфем не виключає варіативність результату. Так, слово «*weichen*» розкладається на морфеми: «*weich*» та «*en*», а також «*weiche*» та «*n*». З цих компонентів синтезується означення (*die* ***weichen*** *Knie* «гнучкі коліна»), дієслово (*er soll* ***weichen*** «він повинен поступитися») або іменник (*für etw. die* ***Weichen*** *stellen* «визначти напрямок або шляхи розвитку»), тому граматичний клас слова та його синтаксична функція в реченні ідентифікуються з урахуванням позиції відносно дієслова. Крім того, процедури аналізу та синтезу дозволяють контролювати синтаксичні узгодження членів речення.

Інтегровані модулі перевірки правопису виявляють і візуалізують зайві пробіли, зайві чи відсутні знаки пунктуації, старий або новий правопис лексичних одиниць, друковані помилки, неконсистентне написання слів, до яких відносимо слова з варіативним правописом, напр.: *Wasserstoff-Fahrzeug, Wasserstofffahrzeugen, Wasserstoff-Autos* або: *Fehlertyp, Fehlerklasse, Fehlerkategorie*. На консистентність перевіряються всі реєстрові одиниці термінологічної бази.

Результат ідентифікації неконсистентних слів унаочнює рис. 1.4:

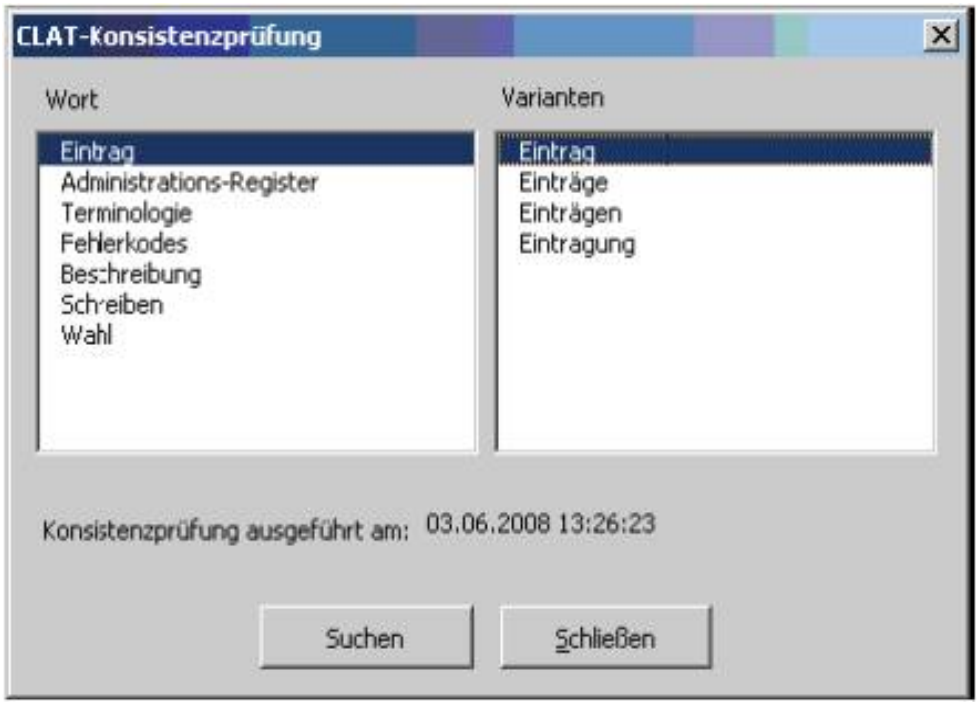

Рис. 1.4. Вікно контролю за консистентним уживанням слова

Ідентифікувати відхилення від усталених норм правопису корпоративної мови або частотні помилки нормативного правопису допомагає ще одна функція –

Falschschreibung «помилка правопису». Адміністратор вносить в UMMT правильно та неправильно написані слова, на ґрунті яких програма в процесі перевірки виявляє неправильний правопис слів і пропонує замінити їх на правильні варіанти написання, напр.: *EFGBiogran* (неправильно) та *EFG-Biogran* (правильно).

Ресурс проекту *Stilregeln* «стилістичні правила» відкриває перелік всіх передбачених програмою стилістичних правил, які можуть «включатись» /«виключатись» з огляду на специфіку проекту.

Цю опцію демонструє рис. 1.5:

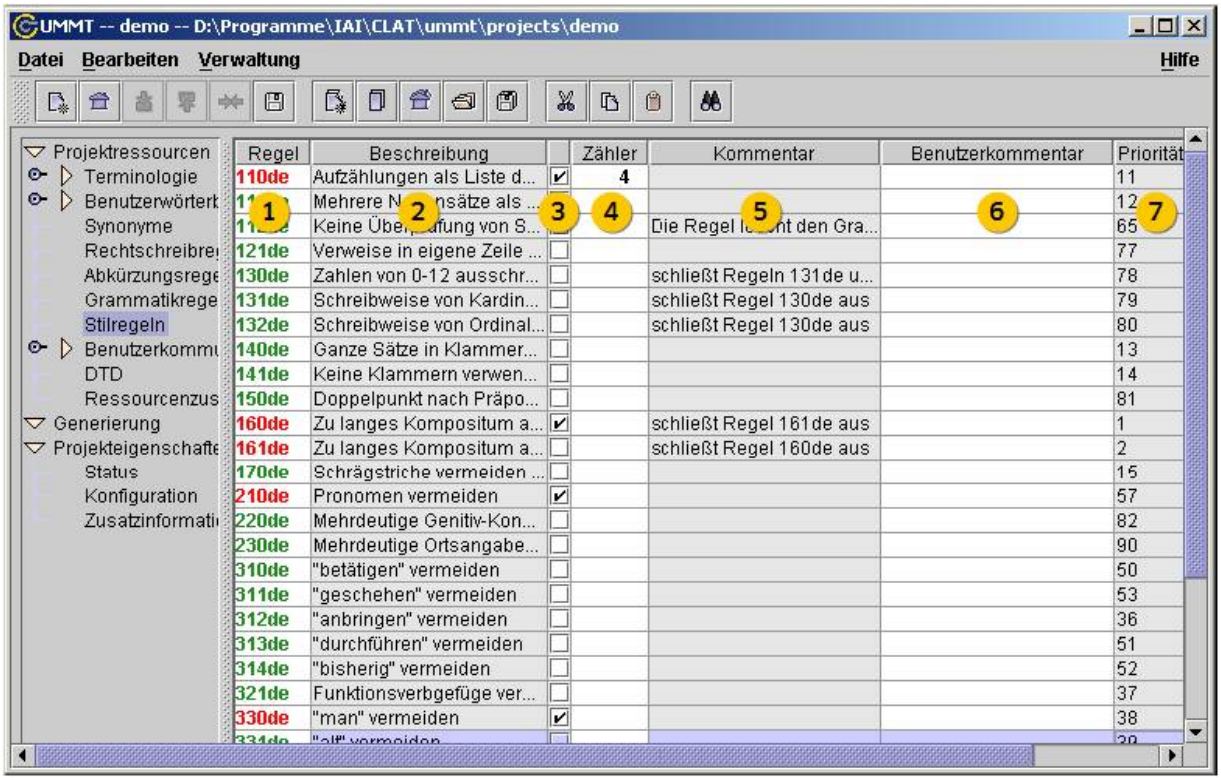

Рис. 1.5. Стилістичні правила CLAT

Правила, марковані червоним кольором, уважаються пріоритетними й активовані за замовчуванням, напр.: уникання неозначено-особового займенника «man», займенників, багатоконституентних композитів та ін. Зелене маркування вказує на додаткові правила, які можуть активізуватись за потреби, напр.: специфічні правила правопису для кількісних і порядкових числівників, уживання синтаксично переобтяжених речень, уживання дужок, двокрапки, косої риски, уживання неоднозначних обставин місця, уживання комплексних прийменникових

словосполучень у родовому відмінку тощо. Маркування синім кольором виокремлює специфічні правила, які програмуються розробником на замовлення та доповнюють загальні правила стилю.

Для роботи з текстами користуються клієнтами. Так, Java CLAT-Client призначений для роботи зі структурованими документами. Робоча область Java CLAT-Client складається з полів: 1) рядок меню; 2) меню символів; 3) зона навігації в тексті; 4) вікно виведення сегмента тексту, який містить помилку за категоріями «Термінологія», «Правопис», «Граматика», «Стиль»; 5) код помилки, інструкція щодо виправлення помилки та правило з граматики Дудена, яке доводить присутність помилки; 6) зона редагування сегмента тексту, який містить помилку; 7) сфера переходів. Робочу область CLAT-Client унаочнює рис. 1.6:

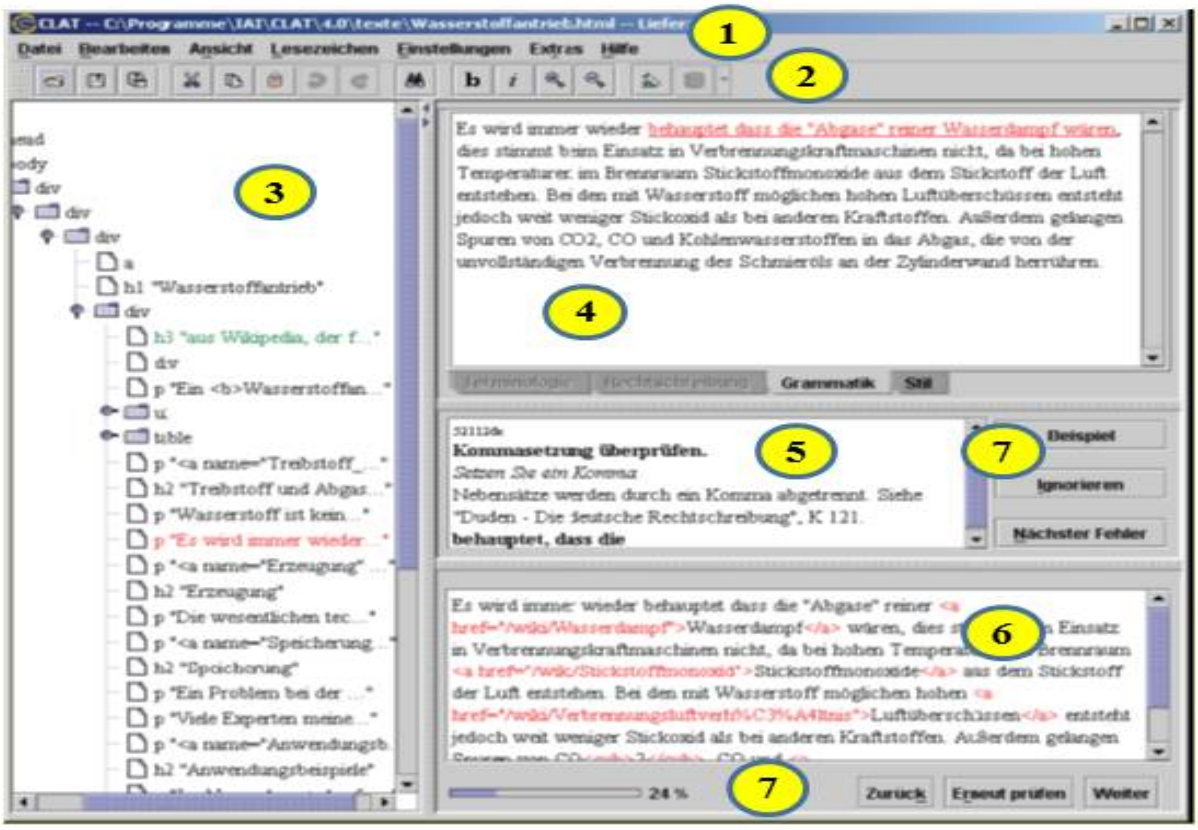

Рис. 1.6. Робоча область Java CLAT-Client

Навігатор CLAT-In інтегрується в ОП Windows та дозволяє контролювати документи з редактора Word за аналогічними опціями: 1) *Fehler* «помилка» – виводить категорію та ідентифікаційний номер помилки (термінологія, правопис, граматика, стиль, напр.: 52111 de); 2) *Fehlermeldungsbereich* «повідомлення про

помилку» – вказує на аспект, який необхідно перевірити, виводить інструкцію щодо виправлення помилки та опорне правило, яке доводить присутність помилки; 3) *Beispiel* «правило» – виводить типову помилку відповідної категорії та приклад її виправлення, який слугує для технічного автора підказкою для переформулювання відповідного фрагмента тексту; 4) *Korrektur* «коректура» – показує вибраний користувачем варіант виправлення помилки; 5) *Vorschläge* «пропозиція» – показує варіанти виправлення помилки, запропоновані CLAT; 6) *Zusatzinformation* «додаткова інформація» – вікно, у яке через опцію «налаштування» можна виводити додаткову інформацію, напр.: належність терміна до певного концепту; 7) *Absatz erneut prüfen* «повторна перевірка абзацу» – після виправлення ідентифікованих програмою помилок запускає опцію повторної перевірки абзацу; 8) *Zurück* «назад» та *Weiter* «далі» – навігація по помилках обраної категорії. Опції навігатора унаочнює рис. 1.7:

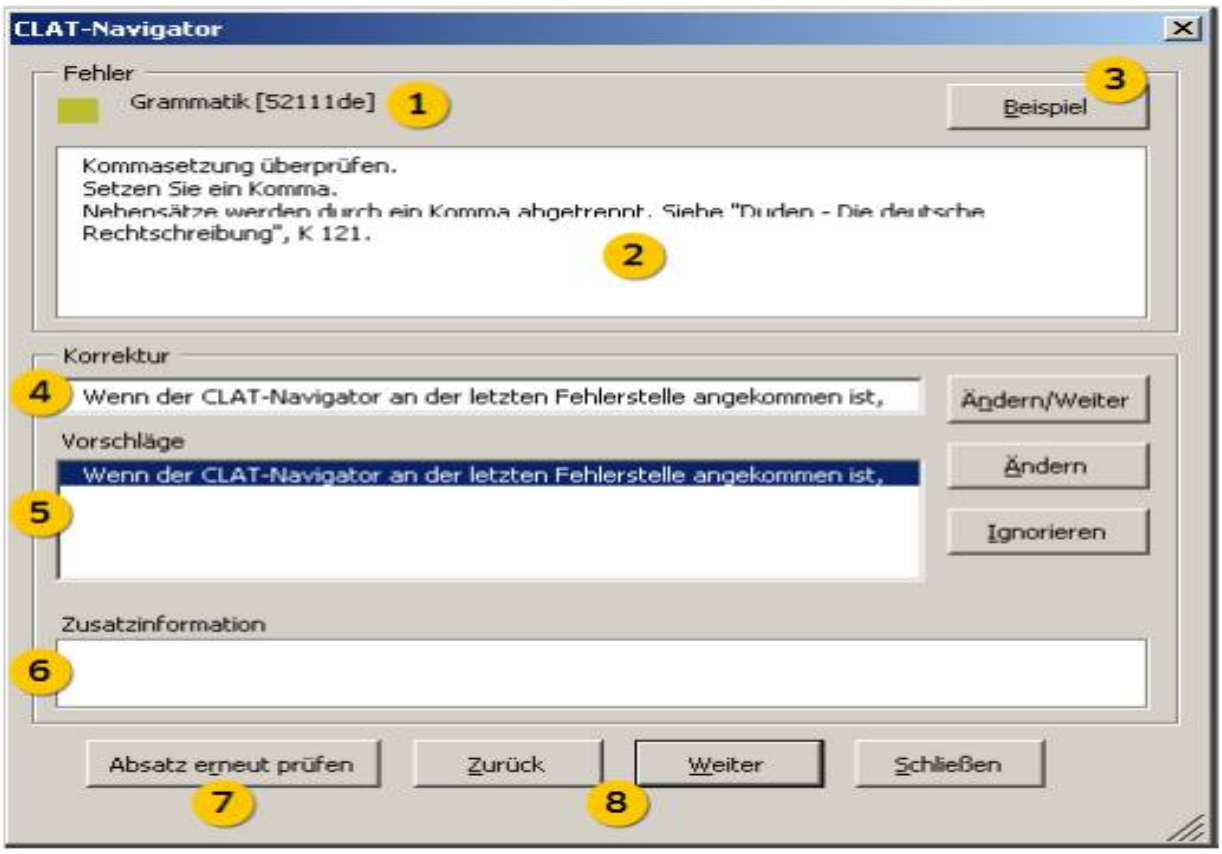

Рис. 1.7. CLAT-навігатор для CLAT-In для Word

Результати перевірки виводяться в окремому вікні та візуалізують: 1) *Status* «статус» – статус процесу перевірки документа; 2) *Letzte Prüfungen* «остання

перевірка» – час і дату останньої перевірки документа; 3) *UMMT-Projekt* «проект UMMT» – назву проекту, на ґрунті якого здійснювалась перевірка документа; 4) *Prüfeinstellungen und -statistik* «налаштування перевірки й статистика» – категорії помилок, які перевірялись, та статистику виявлених програмою помилок з огляду на обрані категорії (правопис, граматика, скорочення, термінологія, стиль, консистентність, «кандидати у терміни»). Вікно статистики помилок демонструє рис. 1.8:

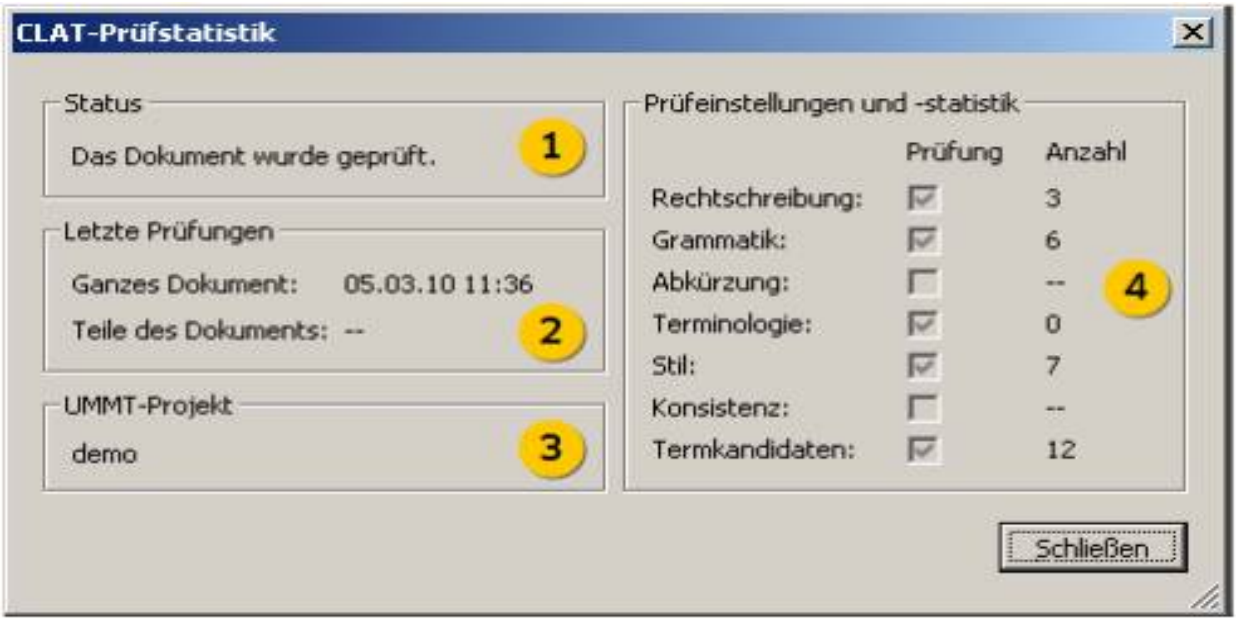

Рис. 1.8. Статистика помилок у CLAT-In for Word

Незважаючи на всі переваги системи, необхідно наголосити, що жодна сучасна програма, у тому числі й CLAT, не здатна перевіряти текст за такими критеріями, як:

- релевантність слів, що уживаються в заголовках та підзаголовках;
- якість формулювання вступу;
- зрозумілість і правильність виокремлення ключових слів з огляду на цільову аудиторію реципієнтів;
- однозначність, імпліцитність / експліцитність, консистентність і повноту формулювання посилань, а також правильність формулювання посилань з урахуванням типів текстів та інструментів їхнього створення;
- якість тексту з огляду на структурованість, повноту, логічність, когерентність викладення матеріалу;

- гомогенність і монотипію формулювання ідентичної інформації, особливо у заголовках і підзаголовках;
- редундантність формулювання;
- влучність формулювання;
- описки, які є орфографічно правильними словами німецької мови, але помилковими з огляду на зміст і контекст висловлювання;
- коректність уживання засобів вираження модальності;
- повноту та релевантність термінів у глосарії, предметному чи системному покажчиках, а також їхнього тлумачення з огляду на розуміння глобального змісту тексту;
- однозначне формулювання ознак та якостей;
- коректність уживання форичних і дейктичних мовних одиниць.

Таким чином, ці програми-коректори не можуть розв’язати всіх проблем технічного автора, але дозволяють оптимізувати роботу над продукуванням уніфікованого та консистентного тексту мовою оригіналу. Обрані автором синтаксичні, морфологічні та стилістичні обмеження запобігають варіативності формулювання, що, у свою чергу, збільшує ймовірність повторного використання сегментів тексту та скорочує витрати і час на створення нового тексту. На другому етапі тексти мовою оригіналу перекладаються іноземними мовами в системах пам’яті перекладу.

Розглянемо, яким чином машинний переклад (МП) вписується в наше уявлення про переведення. Як це не парадоксально, але на даний момент з практичної точки зору машинний переклад залишається процесом людської діяльності.

Термін «машинний переклад» багатозначний. За довгу історію використання він придбав безліч інтерпретацій. Спочатку цей термін мав на увазі тільки автоматичні системи, що працюють без участі людини [43]. Європейська асоціація машинного перекладу дала наступне визначення: «використання комп'ютера для перекладу тексту з однієї природної мови на іншу мову» [Сайт Європейської асоціації машинного перекладу EAMT]. А Міжнародна асоціація машинного перекладу

(IAMT) визначає машинний переклад як «одноразовий введення повного пропозиції і генерування відповідного йому повного пропозиції» [34]. Жодне з цих визначень не передбачає втручання людини.

Академічні вчені та дослідники досі розходяться в поглядах на визначення машинного перекладу щодо участі людини в цьому процесі. В даний момент цей термін продовжує використовуватися для позначення повністю автоматизованих систем нехай навіть і за участю людини [37].

Машинний переклад - це виконується на комп'ютері дію з перетворення тексту на одному природною мовою в еквівалентний за змістом текст на іншій мові, а також результат такої дії [45].

Тлумачний перекладознавчий словник Л. Л. Нелюбина визначає машинний переклад наступним чином:

1. Автоматичний переклад тексту на основі заданої програми, здійснюваної ЕОМ.
2. Галузь мовознавства, що розробляє теорію такого перекладу на основі докорінного перегляду основних положень і методів лінгвістики.
3. Автоматизована обробка інформації в умовах двомовної ситуації - передача тексту з одного людського (природного) мови на іншу.
4. Переклад з використанням машин (ЕОМ, комп'ютера).
5. Загальний процес переробки інформації в умовах двомовної ситуації на будь-якому етапі використання (і розвитку) технічних засобів.
6. Процес перекладу тексту з однієї мови (природного або штучного) на інший (природний або штучний), здійснюваний на електронного цифрового обчислювальної машині [18].

Класифікацію систем машинного перекладу можна зробити за різними підставами [4]. Наприклад, можна виділяти системи:

1. за кількістю мов (бінарні, здійснюють переклад в одній парі мов, і багатомовні, працюють з кількома мовами).

2. спрямованостю перекладу (односпрямовані і у багатьох напрямах, якщо мову перекладу і мова-джерело можуть мінятися місцями залежно від вимог користувача).

Залежно від того, яку роль відіграє людина в процесі МП, іншими словами, за ступенем автоматизації, зазвичай виділяють три типи систем машинного перекладу:

1. Повністю автоматичні системи машинного перекладу.

2. МП-системи, машинний переклад за участю людини.

3. ТМ-системи, переклад здійснюється людиною, при використанні комп'ютера.

Повністю автоматичні системи машинного перекладу є скоріше нездійсненною мрією, ніж реальною ідей. Всі системи машинного перекладу (МП-системи) працюють за участю людини в тій чи іншій мірі. Щоб комп'ютер міг перевести текст, йому потрібна допомога предредактора, який тим чи іншим чином попередньо обробляє підлягає перекладу текст, інтерредактора, який бере участь в процесі перекладу, і постредактора, який виправляє помилки і недоліки в перекладеному машиною тексті [20].

ТМ-системи іноді називають ще «пам'яттю перекладів». Вони є скоріше просто зручним інструментом, ніж елементом автоматизації.

Інший варіант класифікації систем МП, що прийшов з області корпусної лінгвістики, - це поділ на підходи, в яких використовуються паралельні корпусу і, відповідно ті, в яких вони не використовуються. Системи, що використовують корпусу, далі діляться залежно від основної стратегії перекладу - на системи, засновані на прикладах (EBMT), і статистичні системи (SMT) [32].

Найпростіший і найпоширеніший варіант класифікації - це поділ на два основних типи систем МП [16]:

- засновані на правилах (rule-based machine translation, RBMT)
- статистичні

Окремо стоять гібридні системи, які покликані поєднувати в собі найкращі риси систем, заснованих на правилах, і статистичних систем.

Пам'ять перекладів (Translation Memory)

Технологія пам'яті перекладів (Translation Memory або TM) використовує правила перекладу і порівнює вхідний документ з текстами з постійно поповнюється бази перекладів. Знаходячи збіги, програма пропонує раніше схвалений варіант [9].

У процесі перекладу зберігається вихідний сегмент тексту (пропозиція) і його переклад; якщо подібний вихідного сегмент виявляється, він відображається разом з перекладом і зазначенням збігу; потім перекладач приймає рішення (редагувати, відхилити або прийняти переклад), результат якого зберігається системою.

Системи, засновані на правилах (класичні системи)

Технологія цього перекладу полягає в застосуванні алгоритмів, відповідно до яких програма аналізує текст і на основі проведеного аналізу синтезує варіант перекладу.

Вважається, що робота такого машинного перекладача схожа на процес мислення людини [19].

Стандартний алгоритм дій над вхідним пропозицією в такій системі наступний: – морфологічний аналіз – пошук частин мови, визначення вхідних словоформ (роду, числа, відмінка, відмінювання); – пошук ідіом, фразеологізмів для даної предметної області і виключення їх з подальшого аналізу; – синтаксичний аналіз – розбір структури, перебування членів пропозиції – підлягає, присудка, доповнення, обстоятельства; –лексичний аналіз – відділення однозначних вхідних слів (лексем) від багатозначних (що мають кілька перекладних еквівалентів); – граматичний аналіз – доопределение граматичної інформації з урахуванням даних вихідного мови; – синтез вихідного пропозиції (переведення) [9].

У системах, заснованих на правилах (RBMT), можна виділити два основних підтипи: трансферні і системи-інтерлінгва.

Трансферні системи машинного перекладу поширені більш широко, ніж системи-інтерлінгва. Вони працюють за такими принципами: проводиться морфологічний, лексичний і семантико-синтаксичний аналіз пропозиції на мові оригіналу, створюється синтаксично-семантичне дерево розбору вхідного пропозиції, потім проводиться так званий «переклад». Перетворення структури вхідного

пропозиції відповідно до формальних вимог мови перекладу. На заключному етапі синтезу формується кінцевий пропозицію на мові перекладу.

В основі систем-інтерлінгва лежить теорія про те, що будь-яка пропозиція будь-якої мови можна перетворити в його смислове подання на універсальному метамові. Далі, використовуючи отримане смислове уявлення, можна синтезувати пропозицію на мові перекладу. Будь-який текст можна перетворити в сенс, і будь-який сенс в текст, використовуючи ряд правил і семантичний словник. Інтерлінгва вимагають дуже довгої розробки і створення величезних баз знань про мову.

Системи, засновані на правилах, мають ряд загальних характеристик. Всі вони включають в себе словники і формальні граматики. Набори правил морфологічного, семантичного і синтаксичного аналізу мови. З точки зору розробки і використання, такі системи мають ряд переваг і недоліків.

До переваг таких систем можна віднести високу якість, стабільність і передбачуваність машинного перекладу.

Недоліки таких систем включають високу вартість розробки і підтримки лінгвістичних алгоритмів і словників, а також велика кількість часу, необхідне для лексичної настройки системи для окремого клієнта або нової предметної області. Крім того, при високій точності заснований на правилах переклад має певний «машинним» акцентом». Часто виглядає не природно.

Існує також і проблема наростаючої складності. Описати мову в усій його повноті – дуже важке завдання, за рахунок того, що кожен наступний рівень мови виявляється на порядок складніше попереднього, і за рамками опису завжди залишаються деякі лінгвістичні явища.

Сучасні RBMT-системи зазвичай включають в себе загально-тематичні словники (об'ємом від декількох десятків до декількох сотень тисяч статей) і спеціалізовані словники за окремими тематиками (об'ємом до декількох десятків тисяч статей). Продуктивність RBMT – систем машинного перекладу залежить від різних параметрів (серед яких кількість і складність граматичних правил, обсяг і кількість використовуваних словників) і зазвичай варіюється від декількох слів до декількох сотень слів в секунду.

Статистичний МП спирається на припущення, що сказавши щось одного разу, людина з певною ймовірністю повторить це знову [8].

Підхід, який використовується в статистичному МП, полягає в аналізі колосального масиву паралельних текстів. За допомогою цього двомовного паралельного корпусу виявляються пари фраз на двох мовах, які несуть один сенс. При цьому використання якихось додаткових граматичних правил не передбачається [9].

Завдання машинного перекладу в цьому випадку на загальному рівні може бути сформульована як задача максимізації умовної ймовірності P (e | f), що позначає умовну ймовірність пропозиції на мові E при заданому пропозиції на мові F, e∈E, f∈F.

Для виконання цього завдання можна використовувати теорему Байеса. Формула Байєса або теорема Байеса – одна з основних теорем елементарної теорії ймовірностей. Вона дозволяє визначити ймовірність чого-небудь, якої-небудь події за умови, що сталося інше статистично взаємозалежні з ним подія.

Тоді, застосовуючи теорему Байеса, можна записати:

$$P(e \mid f) = \frac{P(e)P(f \mid e)}{P(f)}$$

де

P (e) – апріорна ймовірність гіпотези e;

P (e | f) – ймовірність гіпотези e при настанні події f;

P (f | e) – ймовірність настання події f при істинності гіпотези e;

P (f) – повна ймовірність настання події f

Для максимізації умовної ймовірності зліва потрібно максимізувати величину справа. Наступне рівняння називають фундаментальним рівнянням машинного перекладу [8]:

$$\max_e P(e \mid f) = \arg\max_e P(e)P(f \mid e),\ e \in E, f \in F$$

У цьому рівнянні ймовірність P (e | f) називається моделлю перекладу, а P (e) - мовної моделлю. Побудова цих моделей є частиною навчання статистичної системи МП.

Для використання системи статистичного МП її потрібно спочатку навчити. Процес навчання має на увазі створення двох моделей: статистичної моделі перекладу на підставі паралельного корпусу і статистичної моделі приймаючої мови на основі (часто набагато більшого) одномовного корпусу [27].

Модель перекладу будується по двомовного вирівняні корпусу, тобто такого корпусу, де кожне речення на мові F має переклад на мові E. Інша назва такого корпусу – паралельний корпус.

Побудова такого корпусу є окремою науковою задачею, а отримання паралельних текстів в автоматичному режимі – також і практичної (наприклад, сканування мережі Інтернет в пошуках сторінок, які переводять один одного). Отримання двомовного корпусу на практиці зводиться до аналізу форматів оцифрованих книг-перекладів один одного, а також до індексування Інтернету з метою отримання паралельних сторінок. У цьому випадку можливе застосування різних евристик з розпізнаванням мови і пошуку шаблонів в URL адресах, подібних URL / en і URL / ni. Більш якісним і, відповідно, дорогим способом отримання паралельного корпусу є ручна розмітка. Одним з найбільш популярних джерел паралельних корпусів для пар європейських мов є корпус Europarl [9].

Модель перекладу становить двомовний словник, де для кожного можливого перекладу конкретної одиниці мови-джерела вказана ймовірність такого перекладу [32].

Така модель відрізняється від звичайного словника, де присутні тільки правильні переклади; в цій моделі будуть присутні і малоймовірні переклади. Так, кращим перекладом буде вважатися найвірогідніший, при цьому «кращий» не означає повністю правильний.

Мовна модель створює базу даних типових ланцюжків слів (послідовностей словоформ) в приймаючому мовою (зазвичай від 1 до 7 слів), для кожної з яких вказується ймовірність появи.

Після того, як система була навчена, можна починати процес декодування, тобто безпосередньо використовувати систему для перекладу. Коли система отримує запит від користувача, перекладацька модель генерує можливі варіанти перекладу, а мовна модель вибирає такий переклад, який найбільше нагадує текст, написаний на природній мові.

Коли говорять про статистичному МП, зазвичай мають на увазі фразові перекладачі (Phrase-based translation – PBT).

До появи фразових перекладачів, стандартом вважалися системи послівного перекладу (Word-based translation – WBT). У таких системах кожне слово перекладається окремо, в тому порядку, в якому вони зустрічаються в тексті, без урахування синтаксичних і логічних зв'язків. Поява фразових перекладачів, дозволило враховувати ланцюжка словоформ різної довжини. В системі фразового перекладу вхідний пропозицію ділиться на сегменти, (фрази, ланцюжки словоформ, n-грами) які переводяться окремо. Фраза може складатися з одного і більше слів [33].

Така система дозволяє легко вирішити проблему, коли в приймаючому мовою і мовою джерелі для деяких слів немає точних відповідностей.

У системах послівного перекладу для вирішення цих проблем доводиться водити нові складні стратегії, такі як наприклад нульові слова [32].

Система фразового перекладу має такі переваги:

1. Дозволяє вирішувати лексичну неоднозначність при перекладі полісемантичний слів, з огляду на додаткову контекстуальную інформацію.
2. При збільшенні кількості тренувальної інформації, інформації в тренувальному корпусі, система може вчити все довші фрази. Таким чином, фразові перекладачі використовують тренувальні дані більш ефективно [37].

Багато років фразові системи перекладу показують кращі результати в області МП. В першу чергу, це пов'язано з наявністю величезних паралельних корпусів. Але необхідність використання такого великого обсягу даних може бути проблематична при роботі з мовами, для яких цих даних просто немає. Статистичний переклад як підхід має і інші внутрішні обмеження [44].

Засновані проблеми статистичного перекладу пов'язані з використанням обмеженою лінгвістичної абстракції (limited linguistic abstraction), труднощам перекладу певних конструкцій, як, наприклад, отримання правильного порядку слів при перекладі між мовами різних типів або збереження семантичного єдності в вихідному тексті [44].

До мінусів статистичних систем МП також можна віднести велику кількість граматичних помилок. Окремі словосполучення при статистичному перекладі виходять більш точними і витонченими, але граматика кульгає: іноді пропозиції настільки неузгоджені, що неможливо зрозуміти їх зміст [9].

Іншою проблемою є необхідність наявності представницьких паралельних корпусів великого обсягу.

2. ПРОЕКТУВАННЯ СИСТЕМИ ПОШУКУ СЕМАНТИЧНО-ІДЕНТИЧНИХ ФРАГМЕНТІВ В РІЗНОМОВНИХ ТЕКСТАХ

2.1. Вибір і обґрунтування системи пошуку семантично-ідентичних фрагментів в різномовних текстах

Коли ми говоримо про якість перекладі взагалі, важливо розуміти, що до перекладу, виконаному людиною, будуть пред'являтися значно вищі вимоги. Так, при експертній оцінці перекладу, виконаного людиною, розглядаються такі деталі як прагматика, відповідність перекладу історичного і культурного контексту, стилістика і інші моменти, що стосуються створення сприятливого враження у читача. При перекладі певної лексики можуть розглядатися навіть відтінки значень деяких слів. Розглядати в такому ж ключі машинний переклад представляється неможливим, хоча б з тієї причини, що для усвідомлення метаінформації, яку враховує перекладач при своїй роботі, програма машинного перекладу повинна володіти штучним інтелектом.

Оцінка якості МП є складним завданням, вже хоча б тому, що для вихідного тексту може існувати безліч різних правильних перекладів.

Для оцінки роботи систем МП використовуються наступні методи:

•експертна оцінка

• автоматичні методи

• оцінка з точки зору конкретного завдання

Ми не станемо докладно розбирати оцінку з точки зору конкретного завдання, з тієї причини, що вона повністю залежить від цілей дослідження. В цьому випадку, можуть розглядатися такі питання, як, наприклад, скільки часу йде на постредагування тексту або наскільки точно передається інформація при перекладі [37].

2.2. Постановка задачі моделювання, обґрунтування припущень і розробку базової моделі, аналіз адекватності розроблених моделей

Експертна оцінка

Іноді для оцінки якості перекладу, перевага надається тексту з вузької спеціальної області, і дослідник сам здійснює оцінку якості перекладу, зіставляючи результати роботи декількох систем [14].

Стандартна процедура оцінки має на увазі більше одного експерта.

Експерти проводять суб'єктивну оцінку роботи системи за двома параметрами: адекватність (adequacy) і гладкість тексту (fluency). Для цього їм надають результати роботи системи МП, вихідний текст і / або еталон перекладу. Еталон перекладу часто присутня в тому випадку, якщо експерт не володіє приймають або вихідним мовою. Адекватність і гладкість тексту оцінюються за шкалою від одного до п'яти. Адекватність в даному випадку означає правильну передачу сенсу вихідного тексту, а гладкість тексту демонструє відповідність перекладу нормам приймаючої мови, правильність з точки зору граматики [37].

Одна з проблем, що виникають при такій системі оцінювання, це незгода між експертами. Ця проблема вирішується при використанні коефіцієнта Каппа:

$$K = \frac{p(A) - p(E)}{1 - p(E)}$$

де p (A) - частка випадків, коли експерти дали однакову оцінку, а p (E) - ймовірність того, що експерти випадково дадуть однакову оцінку. Коефіцієнт Каппа дорівнює одиниці буде означати повну згоду експертів [47].

Існує також ранговая система оцінки, коли переклад системи МП попарно порівнюється з перекладами інших систем в термінах «краще» (один з перекладів явно перевершує інший за якістю), «гірше» і «еквівалентно» (переклади принципово не відрізняються за якістю). У цьому випадку експерти, як правило, більш послідовні

в своїх оцінках. Для досягнення неупередженості експерти зазвичай не знають, результати роботи якої системи вони оцінюють.

Автоматична оцінка

Інструменти, що використовуються для автоматичної оцінки МП, в ідеалі, повинні відповідати таким критеріям: низька вартість роботи, інтуїтивно зрозумілі і значущі результати, сталість результатів при повторному використанні і, нарешті, правильність оцінки систем, які працюють краще. Враховується також швидкість роботи, можливість індивідуальної настройки під інтереси користувача і обсяг пам'яті, який потрібно системі [37].

Завдання таких інструментів це, при наявності еталонного перекладу і перекладу, здійсненого МП, порівняти їх і обчислити, наскільки вони схожі.

Для автоматичної оцінки роботи машинних перекладачів часто використовуються показник Word Error Rate або WER, метрики BLEU і NIST. Ці інструменти дозволяють успішно порівнювати роботу різних систем МП і оцінювати поліпшення в роботі конкретної системи [48]. Використовуються також метрики точність (precision), повнота (recall) і F-міра [37].

Розглянемо докладніше принципи їх роботи.

Word Error Rate, або зважена відстань Левенштейна, дозволяє вимірювати відстань між машинним і зразковим перекладом так само, як ми вимірюємо відстань між словниковим словом і словом з помилкою (вважаючи символами не букви, а цілі слова) [МП: огляд методів]. По суті WER вимірює мінімальну кількість змін, які необхідно зробити, щоб з результату роботи МП отримати еталонний переклад [37]. При цьому WER може враховувати різні варіанти еталонного перекладу з різним порядком слів [48].

За формулою зваженого відстані Левенштейна:

$$\mathrm{WER} = \frac{substitutions + insertions + deletions}{reference\text{-}length}$$

де

заміна (substitutions): необхідність заміни одного слова іншим;

вставка (insertions): необхідність додавання слова;

видалення (deletions): необхідність видалення слова;

довжина еталонного перекладу (reference-length).

У випадку з WER, чим менше відстань Левенштейна, тим краще оцінюється робота системи.

Метрика BLEU (Bilingual Evaluation Understudy) на даний момент найпопулярніша в сучасній оцінці МП. Дозволяє враховувати не тільки точність перекладу окремих слів, але і ланцюжків слів (n-грами) [МП: огляд методів].

Метрика BLEU була розроблена співробітниками компанії IBM і є однією з найбільш простих у використанні метрик оцінки машинного перекладу. Алгоритм BLEU оцінює якість перекладу за шкалою від 0 до 100 на підставі порівняння машинного перекладу з людським і пошуку спільних слів і фраз. Основна ідея розробників метрики полягає в тому, що чим краще машинний переклад, тим більше він повинен бути схожий на людський [16].

Варіант метрики BLUE з обмеженням до 4-грам виглядає наступним чином:

$$\text{Bleu} - 4 = min(1, \frac{\text{output} - \text{length}}{\text{reference} - \text{length}}) \sqrt[4]{\prod_{i=1}^{4} precision_i}$$

де

precision - відношення кількості коректних i-грам до загальної кількості i-грам в перекладі;

output - lenght - довжина перекладу, який оцінює метрика;

reference - length - довжина еталлоного перекладу;

Найкраще така метрика працює не на рівні пропозицій, а на рівні великого тексту. На маленькому обсязі тексту метрика часто обнуляється через відсутність співпадаючих 4-грам і не функціонує належним чином. Існують також допрацьовані варіанти метрики, які підходять для порівняння на рівні пропозиції.

Метрика NIST була розроблена на основі BLEU, але має одне фундаментальне відмінність. Якщо для отримання високої оцінки BLEU важливіше правильний порядок слів, то NIST вище оцінює правильний вибір лексики [48].

Для використання метрик BLUE і NIST потрібно корпус пропозицій мовою оригіналу і різні еталонні переклади цих пропозицій, виконані людиною [48].

Очевидно, що ні метрика BLUE, ні NIST не працюють так само як експертна оцінка. Експерти вище оцінюють граматично вірні переклади, які нагадують тексти на природній мові, тоді як метрики оцінюють тексти в межах 5-грам [48], а значить, не можуть оцінити те, наприклад, як пов'язані між собою пропозиції в перекладеному тексті.

Хоча використання метрик BLEU і NIST, стає все більш популярним, ми не до кінця розуміємо, як саме вони працюють [48], Часто результати їх роботи складно інтерпретувати, а з'ясувати причини помилок, що з'являються в конкретній системі, за допомогою тільки цих заходів неможливо [47].

Те ж стосується і всіх інших інструментів автоматичної оцінки роботи МП в цілому. Сама по собі оцінка роботи системи без якихось додаткових досліджень, не надає корисної інформації, яку можна було б використовувати для подальшого розвитку системи МП. Одним з варіантів такого дослідження є детальний аналіз результатів роботи системи і з'являються при роботі помилок.

2.3. Розробка алгоритму і методики проведення моделювання

Алгоритм роботи програм-перекладачів не ідеальний. Тому текст, отриманий з їх допомогою, потребує ретельного постредагуванні. Протестуємо найбільш популярні з програм-перекладачів: Promt і Google Translate. Основні труднощі, з якими стикаються програми перекладачі, такі:

програма при перекладі не враховує багатозначність слів (табл. 2.1);

незрозумілий контекст слів, т. е. умови вживання слова, що дозволяють уточнити їх значення (табл. 2.2).

Таблиця 2.1

Початковий текст	Машинний переклад	Ручний переклад
Комп'ютер запускає агрегат	The computer starts the unit	The system serves actuate / energize the mechanism.
Комп'ютер запускає лазер	The computer starts the laser.	The computer triggers the laser.
Система запущена	The system is started.	T he system was started.

Таблиця 2.2

Початковий текст	Google Translate
I'm going to make you mine	Я збираюся зробити вам моє ...
Just execute the insta l ler	Просто запустіть установщика

Оскільки програма не завжди враховує контекст слів (різні версії продукту справляються по-різному), з'являються подібні перекладацькі казуси.

Як було сказано вище, алгоритм роботи програми заснований на аналізі синтаксичних і морфологічних форм вихідного тексту. Тому для більш коректного перекладу з флективного мови необхідно не тільки постредагування перекладу, а й пере редагування вихідного тексту, його адаптування для програми. На основі аналізу принципу р а боти додатків машинного перекладу розроблений наступний алгоритм адаптації тексту для наступного перекладу:

Речення повинні мати чітку граматичну структуру; для коректного перекладу на аналітичні мови необхідно дотримуватися прямого порядку слів, тобто підлягають і його група + присудок + доповнення + обставину.

Використовуючи рекомендовані прийоми, можна значно підвищити якість перекладу, хат е жати грубих помилок. Але важливо пам'ятати, що навіть при скрупульозної підготовки вихідного тексту до перекладу програмою, постредагування.

Крім перерахованих вище інструментів сучасного перекладачеві можуть знадобитися і багато інших: графічні редактори, програми для верстки, підготовки до публікації в мережі (Adobe Publisher) і навіть програми інженерної графіки (AutoCad, 3 D Max).

3. РЕАЛІЗАЦІЯ СИСТЕМИ ПОШУКУ СЕМАНТИЧНО-ІДЕНТИЧНИХ ФРАГМЕНТІВ В РІЗНОМОВНИХ ТЕКСТАХ

3.1. Експериментальні дослідження системи визначення ідентичності різномовних текстів

В ході нашого експерименту, нами було проаналізовано 15043 реальних користувальницьких запитів до онлайн-перекладачеву.

Ми використовували класифікацію типів помилок, наведену у другому розділі, вибрали 300 перших запитів з обох списків (2% від загального числа запитів), порахували кількість помилок того чи іншого класу, які ми коротко описали в пункті 2.2., І отримали для нашого корпусу наступні приблизні статистичні дані:

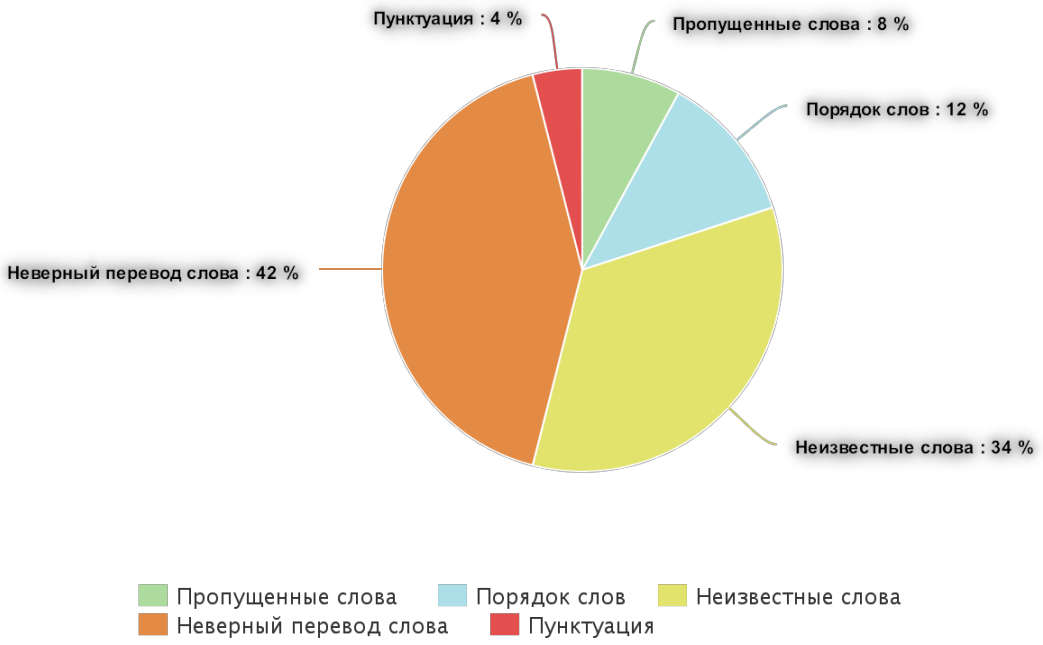

Рис.3.1. Діаграма процентного співвідношення помилок відповідно до класифікації, наведеної в пункті 2.2.

Ми виявили, що найбільше помилок відносяться до категорії «неправильний переклад слова». Приблизно третина з усіх помилок можна віднести до класу

«невідомі слова». З цього ми можемо зробити висновок про недостатню показності корпусу.

За допомогою докладного аналізу помилок, враховуючи теоретичні значення про лінгвістичні особливості мов і принципах роботи статистичного МП, можна виявити причини помилок і запропонувати можливості їх виправлення.

Основні проблеми в роботі інструменту Google Translate викликані недостатньою показністю корпусу, який використовувався для створення перекладацької моделі. Також в корпусі міститися деякі спочатку невірні переклади і помилки, що призводить до повторення тієї ж помилки при роботі системи.

Роботу в деякій мірі порушує присутність іноземних слів в корпусі, що використався для побудови моделі мови.

Відсутність попередньої обробки запиту призводить до того, що користувачі оформляють запити невірно і на виході отримують незадовільний результат.

Система не завжди справляється з перекладом назв, або залишаючи їх в початковій формі, або переводить їх як звичайні слова.

В роботі алгоритму перекладу присутній повторювана систематична помилка, пов'язана із занадто великим вікном перекладу.

Користувачі, які не знайомі з принципами роботи статистичного МП, часто некоректно оформляють запити, що призводить до отримання незадовільних результатів.

ВИСНОВКИ

У даній роботі ми визначили поняття машинного перекладу, описали основні типи систем і методи оцінки МП. На підставі вивченої нами теоретичних даних, описаних в першому і другому розділах, ми проаналізували роботу статистичного онлайн-перекладача Google Translate, детально розібрали помилки, що з'являються при роботі цієї системи, привели нашу власну класифікацію помилок і запропонували способи їх усунення.

У першому розділі ми описали історію розвитку систем МП, сучасний стан цієї галузі, і розглянули три основних сучасних підходу до МП: заснованого на правилах, статистичного і гібридного. Далі, у другому розділі ми описали популярні способи оцінки МП, експертну оцінку і різні метрики. Ми також привели одну з можливих класифікацій помилок, що з'являються в ході роботи систем МП.

У третьому розділі для аналізу роботи онлайн-перекладача Google Translate ми використовували корпус з 15043 реальних запитів користувачів (295 тисяч). Ми привели статистичні дані типів помилок і дали свою власну класифікацію помилок, на основі причин їх появи.

Дослідницькі роботи, що проводяться в області МП, часто можна розділити на дві категорії: написані з точки зору лінгвістики, і написані з точки зору точних обчислювальних наук. Так, роботи, в яких дається оцінка якості перекладу, часто повністю опускають або не враховують принципи роботи програм, які використовуються для цього перекладу. Дослідження, які не враховують даних лінгвістики, надають статистичні дані про кількість і типи помилок, оцінки BLEU або NIST, які складно інтерпретувати. У підсумку, це призводить до того, що причини появи помилок залишаються за межами дослідження. Для поліпшення результатів таких досліджень, фахівці з різних областей повинні більше взаємодіяти.

Для подальшого розвитку систем переказу, які будуть використовуватися реальними користувачами, потрібно розуміти, як і хто в кінцевому підсумку буде ними користуватися. Потрібно враховувати потреби користувача. Так для професійних перекладачів буде корисна можливість вибору між декількома

варіантами перекладу, а для звичайного користувача будуть потрібні різні інструменти, які здійснюють попередню обробку запиту.

Наше власне дослідження запитів продемонструвало, що користувачі переводять тексти різних функціональних стилів, і обмежити тематику або стилістику текстів практично неможливо. Проте, можна стверджувати, що значна частина запитів стосується області комерції і розваг. Ці дані можна враховувати в подальшому при складанні корпусу для перекладацької моделі.

Я вважаю завдання, поставлені в даній роботі, виконаними, а мета - досягнутою.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Аматов А. М. К вопросу машинного перевода: энтропия языковой системы и способы ее преодоления // Вестник ЛГУ им. А.С. Пушкина. 2008. №2 (13) С.71-90.
2. Ахманова О. С. Словарь лингвистических терминов. М., 1969.
3. Бархударов Л. С. Язык и перевод. М.,1975.
4. Беляева Л. Н. Лингвистические автоматы в современных гуманитарных технологиях: Учебное пособие. СПб, 2007.
5. Борисова И. А. К опыту постредактирования на материале англо-русского перевода с помощью автоматических систем Google translate и Prompt // Вестник МГЛУ. 2014. №13 (699) С.53-59.
6. Борисова И. А. Коммуникация между интернет-пользователями — носителями различных языков // Вестник МГЛУ. 2013. №13 (673) С.28-34.
7. Гальперин И. Р. Введение. // Большой англо-русский словарь. М., 1987.
8. Кан, Д. А. Применение теории компьютерной семантики русского языка и статистических методов к построению системы машинного перевода: диссертация кандидата физико-математических наук. Место защиты: Федеральное государственное образовательное учреждение высшего профессионального образования Санкт-Петербургский государственный университет. Санкт-Петербург, 2011.
9. Карасев И. В., Артюшина Е. А. Системы машинного перевода // Успехи современного естествознания. 2011, №7, С.117-118.
10. Колшанский Г. В. Контекстная семантика. М., 1980.
11. Комиссаров В. Н. Современное переводоведение. Учебное пособие. М., 2002.
12. Красных В. В., Изотов А. И. Язык, сознание, коммуникация: Сборник статей. М., 2011.
13. Латышев Л. К. Перевод: проблемы теории, практики и методики преподавания. М., 1988.
14. Максименко О. И., Чинина Д. С. Обзор системы машинного перевода «Google Переводчик» (на примере финского языка). // Science Time, 2014, №5 (5), С.133-139.
15. Марчук Ю. Н. Компьютерная лингвистика: учебное пособие. М., 2007.
16. Молчанов А. Статистические и гибридные методы перевода в технологиях компании ПРОМТ. М., 2013.
17. Найда Ю. К науке переводить // Вопросы теории перевода в зарубежной лингвистике. М.,1978.
18. Нелюбин Л. Л. Толковый переводческий словарь. М., 2011.

19. Новожилова А. А. Машинные системы перевода: качество и возможности использования // Вестник ВолГУ. Серия 2: Языкознание. 2014. №3 С.67-73.
20. Рябцева Н. К. Информационные процессы и машинный перевод. Лингвистический аспект. М., 1986.
21. Слокум Дж. Обзор разработок по машинному переводу. Новое в зарубежной лингвистике. М., 1989.
22. Федоров А. В. Основы общей теории перевода (лингвистические проблемы). М., 2002.
23. Фролов С. В., Паньков Д. А. Проблемы построения машинного перевода. Тамбов, 2008.
24. Шаляпина З. М. Автоматический перевод: Эволюция и современные тенденции // Вопросы языкознания, 1996, №2, С. 105—117.
25. Шевчук, В. Н. Информационные технологии в переводе. Электронные ресурсы переводчика. М., 2013.
26. Baker M. Routledge Encyclopedia of Translation Studies. London & New York, 2001.
27. Brown P. F., Delia Pietra V. J., Delia Pietra S. A., Mercer R. L. The mathematics of statistical machine translation: Parameter estimation // Computational Linguistics, 1993, Vol. 19, №2, P. 263—311.
28. Burukina, I. Translating implicit elements in RBMT. // Translating and the Computer 36, 2014, Asling, P. 182—193.
29. Costa-jussà, M., Fonollosa, J. Latest trends in hybrid machine translation and its applications. // Computer Speech & Language, 2015, №32(1), P. 3-10.
30. Guzmán F., Joty S., Marquez L., Nakov P. Using Discourse Structure Improves Machine Translation Evaluation. // ACL (1), 2014, P. 687-698.
31. Härmävaara H. Trouble sources in Finnish-Estonian RM interaction. Helsinki, 2015.
32. Hearne M., Way A. Statistical Machine Translation: A Guide for Linguists and Translators // Language and Linguistics Compass, 2011, №5, P. 205-226.
33. Heyn M. Integrating MachineTranslation into Translation Memory Systems.// Proceedings of the EAMT Machine TranslationWorkshop, Vienna, Austria, 1996, P. 113—126.
34. Hutchins, 2000a — John Hutchins. Hutchins J. The IAMT Certification Initiative and Defining Translation System Categories // Proceedings of 5th EAMT Workshop, Slovenia, 2000.
35. Hutchins, 2000b — John Hutchins. Petr Petrovich Troyanskii (1894-1950): A forgotten pioneer of mechanical translation. // Machine Translation, vol. 15 no. 3, 2000. P. 187—221.
36. Jehl L. Machine translation for Twitter. Master's thesis. The University of Edinburgh, 2010.

37. Koehn, P. Statistical Machine Translation. Cambridge, UK, 2010.
38. Koponen M., Salmi L. On the correctness of machine translation: A machine translation post-editing task. // The Journal of Specialised Translation, 2015, №23, P. 118—136.
39. Koppel M., Ordan N. Translationese and its dialects. // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Techologies, Portland, Oregon, 2011, P. 1318—1326.
40. Kurokawa D., Goutte C., Isabelle P. Automatic detection of translated text and its impact on machine translation. // Proceedings of MT-Summit XII, 2009, P. 81—88.
41. Lembersky G., Ordan N., Wintner S. Improving statistical machine translation by adapting translation models to translationese. // Computational Linguistics, 2013, №39(4), P. 999—1023.
42. Llitj´os A., Carbonell J., Lavie A. A framework for interactive and automatic refinement of transfer-based machine translation. // Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT), Budapest, Hungary, 2005.
43. Sager J. C. Language Engineering and Translation: Consequenses of Automation. Amsterdam, 1994.
44. Silva J., Rodrigues J., Gomes L., Branco A. Bootstrapping a hybrid deep MT system. Lisbon, 2015.
45. Somers H. L. Intoduction // Computers and Translation: A Translator's Guide. Amsterdam, 2003.
46. Twitto-Shmuel, N., Ordan, N., Wintner, S. Statistical machine translation with automatic identification of translationese. // Proceedings of WMT-2015, 2015
47. Viera A., Garrett J. Understanding interobserver agreement: The Kappa Statistic. // Family Medicine, 2005, №37, P. 360-363.
48. Vilar D., Jia Xu, D'Haro L., Ney H. Error Analysis of Machine Translation Output. In International Conference on Language Resources and Evaluation, pages 697—702, Genoa, Italy, 2006.
49. Zhang Y., Vogel S., Waibel A. Interpreting BLEU/NIST Scores: How Much Improvement do We Need to Have a Better System? // Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-2004), Lisbon, Portugal, 2004.
50. Zhechev V., Genabith V. Seeding Statistical Machine Translation with Translation Memory Output through Tree-Based Structural Alignment. // Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation, Beijing, China, 2010, P. 43—51.

ИНТЕРНЕТ-ИСТОЧНИКИ

51. Сайт Европейской ассоциации машинного перевода EAMT. European Association for Machine Translation EAMT.

URL: http://www.eamt.org/mt.html (дата обращения: 6.01.2016)

52. Сайт Джона Хатчинса.

URL: http://www.hutchinsweb.me.uk/history.htm (дата обращения: 6.01.2016)

53. Лекция о системах МП - Системы автоматического (машинного) перевода текста. История, основные сведения, описание. Лекция №13.

URL: http://itclaim.ru/Education/Course/Lingvistika/Lecture/Lecture13.pdf (дата обращения: 9.11.2015).

54. Онлайн версия доклада ALPAC.

URL: http://www.nap.edu/openbook.php?isbn=ARC000005 (дата обращения: 6.12.2015)

55. МП: обзор методов - Презентация: Математические модели в лингвистике 7. Машинный перевод: обзор методов и оценка качества.

URL: http://lpcs.math.msu.su/~pentus/mfk2015/Lecture07_20151021.pdf (дата обращения: 9.11.2015).

Додаток 1

Система пошуку семантично-ідентичних фрагментів в різномовних текстах

Специфікація

УКР.НТУУ“КПІ”.ТР4174_18Б

Аркушів 2

2019

Позначення	Найменування	Примітки
Документація		
УКР.НТУУ«КПІ ім. Ігоря Сікорського».ТР5174_18Б 81-1	Записка	Пояснювальна записка
Компоненти		
УКР.НТУУ«КПІ».ТР5174_18Б 12-1	Текст програмного модулю	
УКР.НТУУ«КПІ».ТР5174_18Б 13-1	Опис програми	

Додаток 2

Система пошуку семантично-ідентичних фрагментів в різномовних текстах

Текст програмного модуля

УКР.НТУУ“КПІ”.ТР5174_18Б 12-1

Аркушів 5

2019

```
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;
import org.simmetrics.StringMetric;
import org.simmetrics.metrics.StringMetrics;

import java.io.BufferedReader;
import java.io.IOException;
import java.io.InputStreamReader;
import java.net.HttpURLConnection;
import java.net.MalformedURLException;
import java.net.URL;
import java.util.ArrayList;
import java.util.Arrays;
import java.util.List;


/**
 * create a new PlagiarismChecker object for every paragraph to be checked.
 * <p>
 * TODO:
 * Check each string through search engine
 * collect top X results to be searched
 * rip text from HTML of website
 * search text to see if it contains string
 * use algorithm to compare two strings
 * If match is found flag text & record source
 */

public class PlagiarismChecker {


    private ArrayList<PlagData> list = new ArrayList<>();
    public static final String GOOGLE_SEARCH_URL = "https://www.google.com/search";

    public PlagiarismChecker(String paragraph) {
     String[] sentence = paragraph.split("\\.");
        for (int i = 0; i < sentence.length; i++) {
            list.add(new PlagData(sentence[i]));
        }
    }

    /**
     * begins entire plagiarism checking method
     */
    public ArrayList<PlagData> start() {
        for (int i = 0; i < list.size(); i++) {
            try {
                //retrive search results for current sentence.
                Elements results = googleSearch(list.get(i).getSentence());
                if (results == null) {
                    continue;
                }
                // for every result get the link and begin comparison
                for (Element e : results) {
                    String linkHref = e.attr("href");
                    //System.out.println("Link = "+linkHref.substring(6,
linkHref.indexOf("&")));
                    String url = linkHref.substring(6,
linkHref.indexOf("&")).substring(1);
                    comparison(url, list.get(i));
                }
```

```
            } catch (IOException e) {
                e.printStackTrace();
            }
        }
        return list;
    }


    private Elements googleSearch(String sentence) throws IOException {
        String searchSentence = sentence.replace(' ', '+');

        if (searchSentence.length() < 2) {
            return null;
        }

        String searchURL = GOOGLE_SEARCH_URL + "?q=" + searchSentence + "&num=" +
Config.RESULTS;
        Document doc = null;

        try {
            doc = Jsoup.connect(searchURL).userAgent("Mozilla/5.0").get();
            return doc.select("h3.r > a");
        } catch (IOException e) {
            System.out.println("Bad URL");
            return null;
        }

    }


    /**
     * Compares every sentence to the each sentence from the webpage
     * flags sentence if similiarity is to high
     *
     * https://github.com/mpkorstanje/simmetrics
     * string comparison library instructions
     */
    private void comparison(String url, PlagData data) throws IOException {
        //split website content by sentence into arraylist
        ArrayList<String> webContent = retrieveWebContent(url);
        StringMetric metric = StringMetrics.cosineSimilarity();
        //compare each sentence from website with currently selected user sentence
        for (String webSentence : webContent) {
            float similarity = metric.compare(webSentence, data.getSentence());
            if (data.getPlagiatPercent() <= similarity) {
                data.setPlagiatPercent(similarity);
            }

            if (similarity >= Config.SIMILIARITY_LIMIT) {
                flagSentence(url, data);
                return;
            }
        }
    }

    private void flagSentence(String url, PlagData data) {
        data.setFlagged(true);
        data.setSource(url);
    }


    private ArrayList<String> retrieveWebContent(String webAddress) throws
IOException {
```

```
        BufferedReader br = null;
        StringBuilder sb = new StringBuilder();
        // use string builder to increase the performance of repeated string
additions
        try {

            URL url = new URL(webAddress);

            HttpURLConnection connection = (HttpURLConnection) url.openConnection();
            connection.addRequestProperty("User-Agent", "Mozilla/4.76");

            br = new BufferedReader(new
InputStreamReader(connection.getInputStream()));

            String line;
            //read website line by line maintaining formatting
            while ((line = br.readLine()) != null) {
                sb.append(line);
                sb.append(System.lineSeparator());
            }

            String webContent = sb.toString();

            webContent = Jsoup.parse(webContent).text();

            //check if doc has content
            System.out.println("WEBSITE FOUND:\t" + webAddress);

            List<String> temp = Arrays.asList(webContent.split("\\."));
            ArrayList<String> arrayList = new ArrayList<>(temp);
            return arrayList;
        } catch (MalformedURLException e) {
            System.out.println("Malformed URL @ retrieve webcontent");
        } catch (ClassCastException e) {
            System.out.println("class catch @ retrieve webcontent");
        } finally {
            if (br != null) {
                br.close();
            }
        }

        return null;
    }


}

import javax.swing.*;
import javax.swing.border.EmptyBorder;
import java.awt.*;
import java.util.ArrayList;

public class Reports extends JFrame {

    /**
     *
     */
    private static final long serialVersionUID = 1L;
    private JPanel contentPane;
    private static ArrayList<PlagData> plag;
    /**
     * Launch the application.
     */
```

```
    public static void main(String[] args) {
        EventQueue.invokeLater(new Runnable() {
            public void run() {
                try {
                    Reports frame = new Reports();
                    frame.setVisible(true);
                } catch (Exception e) {
                    e.printStackTrace();
                }
            }
        });
    }

    /**
     * Create the frame.
     */
    public Reports() {
        setTitle("Report");
        setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
        setBounds(100, 100, 450, 300);
        contentPane = new JPanel();
        contentPane.setBorder(new EmptyBorder(5, 5, 5, 5));
        setContentPane(contentPane);
        contentPane.setLayout(null);

        JLabel reportLabel = new JLabel("Text"/*fillReport()*/);
        reportLabel.setBounds(10, 11, 300, 60);
        reportLabel.setFont(new Font("Tahoma", Font.PLAIN, 16));
        reportLabel.setVerticalAlignment(SwingConstants.TOP);
        contentPane.add(reportLabel);

        JTextArea commentBox = new JTextArea();
        commentBox.setBounds(10, 100, 412, 150);
        contentPane.add(commentBox);


        JButton btnSave = new JButton("Save");
        btnSave.setBounds(328, 11, 96, 39);
        contentPane.add(btnSave);
    }

    public static String fillReport()
    {
        /*
            This function creates the report which is then displayed in the displayed
label.
            There are html code blocks within the string as to format the text to the
desired look
         */

        String report;
        int totalSentences = 0, detectedSentences = 0;
        double percentage = 0;

        for (PlagData aPlag : plag) {
            totalSentences++;
            if (aPlag.isFlagged() == true) {
                detectedSentences++;
            }
        }

        percentage = ((double)detectedSentences / (double)totalSentences)*100;
```

```
        report = "<html>Total sentences: " + totalSentences + "<br>";
        report += "Plagiarised sentences detected: " + detectedSentences + "<br>";
        report += "Percentage of text plagiarised: " + percentage + "%</html>";

        return report;
    }

    public static void setTextData(ArrayList<PlagData> plagData) { //Setter
        plag = new ArrayList<>(plagData);

    }
}
```

Додаток 3

Система пошуку семантично-ідентичних фрагментів в різномовних текстах

Опис програмного модуля

УКР.НТУУ“КПІ”.ТР5174_18Б 13-1

Аркушів 5

2019

АНОТАЦІЯ

Дипломну роботу виконано на 70 аркушах, вона містить 3 додатки та перелік посилань на використані джерела з 50 найменувань. У роботі наведено 6 рисунків.

Метою роботи було створення системи пошуку семантично-ідентичних фрагментів в різномовних текстах. Програма забезпечує пошук семантично-ідентичних фрагментів в різномовних текстах, забезпечує переклад текстів на англійську мову. Розроблений програмний продукт може бути використаний, наприклад, в організаціях та установах, де часто застосовується перевірка на плагіат.

ЗМІСТ

ВІДОМОСТІ ПРО ПРОГРАМНИЙ МОДУЛЬ

Даний програмний модуль розроблено у середовищі IntelliJ IDEA, , використовуючи типізовану мову програмування Java, не типізовану мову програмування JavaScript, бібліотеку Three.js та деякі додаткові бібліотеки.

Програма призначена для пошуку семантично-ідентичних фрагментів в різномовних текстах.

1.1. Опис логічної структури

Було розроблено багатоплатформний програмний продукт, основною задачею якого є пошук семантично-ідентичних фрагментів в різномовних текстах. При пошуку запозичених фрагментів текстів системи перевіряють на наявність послідовності ідентичних слів та враховують це за плагіат. Цей додаток має значну перевагу перед сучасними та наявними системами, оскільки має змогу знаходити послідовність фрагментів текстів за допомогою пошуку семантично-ідентичних фрагментів.

- надання користувачу легкого та зрозумілого інтерфейсу;
- забезпечення швидкого доступу до файлів з різномовними текстами;
- переклад з різних мов на англійську мову;
- оптимізація взаємодії користувача з системою;
- виділення стоп-слів.

Розроблена програма має коректний механізм пошуку семантично-ідентичних фрагментів в різномовних текстах. Завдяки цьому така система може бути використана в сфері, де потрібна перевірка вхідних текстів на плагіат наукових робіт для визначення унікальності інформації, яку хоче запропонувати автор.

1.2. Вхідні та вихідні дані

Вхідними даними для системи є текстові файли, один з яких оцінюватиметься на відсоток запозиченості, та решта інших, які виступають в ролі ресурсу, з яким буде перевірятися вхідний текст.

Вихідними даними є звіт про пошук семантично-ідентичних фрагментів в різномовних текстах.

ВИКОРИСТАНІ ТЕХНІЧНІ ЗАСОБИ

Програмний модуль було протестовано на лептопі з операційною системою macOS Mojave 10.14.4, який працює на базі процесору 2,6 GHz Intel Core i7 та має 16 Гб оперативної пам’яті. Розроблене програмне забезпечення є кросплатформенним, що дозволяє запускати його на комп’ютерах будь-якої потужності